# A THEORETICAL AND EXPERIMENTAL STUDY OF THE NATURE AND EXTENT OF PREDETERMINATION OF SCORE-SCATTER BY THE TYPE OF TEST-PAPER USED.

A Thesis

presented to

the University of Edinburgh

by David Arnold Walker, M.A., B.Ed..

1937

# C O N T E N T S

PART ONE

## Chapter One. Introduction and non-mathematical summary of the investigation and its results.

### 1. Introduction.

Tests and examinations are used for many different purposes. There is the familiar school class examination, the main object of which is to ascertain how much of the subject taught has been understood and remembered. An examination of a slightly different type is that set for the Leaving Certificate of the Scottish Education Department; although it follows the lines of the ordinary schbol examination, an important point is that it is used to classify the candidates into two categories - those worthy to pass, and those who must be counted as failed. A third type of test is exemplified in the examination set to candidates for posts in the Civil Service; what is important in this test is the order of merit of the leading candidates.

The different purposes of these examinations are reflected in the structure of the papers. The examiners, through much experience of their particular type of work, have built up a technique of construction of papers, and this, we venture to say, is largely of the "rule of thumb" type. It is the purpose of the research reported here to seek the theoretical basis of the design and effect of different types of paper.

2. Definition of Score-scatter.

If the results of the examination are arranged so as to show how many times each score in the possible range was made, this table is called the Score-scatter. For each score x we have tabulated the value of $N_x$, the number of candidates making that score. Where there are many items in a test, it is usual to group the scores; the score-scatter will then show how many candidates scored from 0 -10, from 10 - 20, and so on. Graphically the score-scatter is often shown by histograms.

The examiner is interested in the nature of this distribution. If the examination is of the "pass" type, as exemplified by the Leaving Certificate examination mentioned above, he wishes as clear a demarcation as possible between the "passes" and "fails". When the test is of the Civil Service type, on the other hand, the examiner wishes his test to show as clearly as possible the order of merit of the leading candidates; that is he prefers the score-scatter to show the candidates making high scores well spaced out, no matter whether those further down the scale are bunched together or not. In the case of a school examination, the examiner normally wishes no group to be bunched unduly, though in a special case he may require a test which will show quite definitely which of a group of well-matched pupils is best in the particular subject.

What factors, then, influence the score-scatter ? Obviously the nature of the population tested affects it greatly e.g. a class selected for its high proficiency in a given subject would yield a different score-scatter from that given by an unselected group. An important fact that must be noted is that when the population tested is unselected and reasonably large, the score-scatter always shows a tendency to normality; i.e. the scores are grouped about the mean in the way characteristic of the normal frequency distribution. In certain cases the distribution of ability or knowledge among the candidates may not be of the usual type; in employing the word usual we make no assumptions as to the nature of the usual type. Such a case would arise if the population were "creamed" of the more able 20% say; it is fairly obvious that this would produce a different score-scatter from that produced by the "uncreamed" population. The type of population is a factor over which the examiner has no control, but one to which he must pay attention in his construction of the test.

Another set of factors is connected with the type of paper set. Without any attempt at completeness we may enumerate such possible factors as the number of questions set, the time allowance, any interlinking of items so that a correct answer to one item is impossible unless the previous one has been correctly answered, the general difficulty level of the paper, and the variations of individual items from that level.

From this complexity we select for first consideration those tests for which the time allowance is adequate, where the items are independent, and where each carry unit score when correctly answered. Under these conditions it will be seen that the only tools the examiner can use in attemptimg to provide, or preordain, the required score-scatter are the number of the scoring points, and the difficulty of the various items.

3. Definition of Answer-pattern.

The varying degrees of difficulty of the questions will be reflected in the differing frequencies with which they are correctly answered. If question 1 is answered correctly $n_1$ times by a given set of candidates, while question 2 is answered correctly $n_2$ times, then we can say that question 1 is more or less difficult than question 2 according as $n_1$ is less or greater than $n_2$. If $n_1$ equals $n_2$ both questions are equally difficult. A useful extension of this nomenclature is to let $n_0$ equal the number of candidates. The table of these n's, including $n_0$, for a particular test, we call the Answer-pattern of the test. ( This term was, I believe, first proposed by Professor Godfrey H. Thomson.)

Note that a test has not a unique answer-pattern; the answer-pattern depends on the character of the group of candidates tested as well as on the difficulty of the questions.

In the subsequent work it may be assumed, unless other-

-wise stated, that the items of each test have been placed in ascending order of difficulty, i.e.,

$n_o \geq n_1 \geq n_2 \geq \ldots\ldots\ldots \geq n_m$.

Examples of answer-patterns and score-scatters will be found in subsequent chapters.

The answer-pattern is a distribution of which an examiner has a reasonable chance of knowing something before the test is given in its final form. Often a trial form of the test is given to another group of candidates of the same ability level to ensure that no ambiguities or unexpected difficulties are present. From this form the final form of the test is constructed by eliminating the unsatisfactory items, and perhaps by other slight changes in the format of the paper. After these changes the answer-pattern of the remaining items is still known. We cannot assume that the items left are entirely unaffected by the changes in the paper, but this problem will be discussed in detail in a later chapter.

## 4. The Relation of Answer-pattern and Score-scatter under Unig and Hig conditions.

It is the thesis of this paper that the answer-pattern exerts quite a strong measure of control over the nature of the score-scatter. From a knowledge of the answer-pattern it is possible to predict with a fair degree of accuracy the score-scatter that will be produced.

From the answer-pattern there can be obtained directly the distribution

$n_0-n_1$ , $n_1-n_2$ , $n_2-n_3$ ,....., $n_{m-1}-n_m$ , $n_m$ ,

which we may call the answer-pattern-differential.

In the particular case where every candidate's score is made up of answers to the easiest possible items, the score-scatter is identical with the answer-pattern-differential. This state of affairs is termed 'unig' since any score is compiled in a unique manner. What is even more important is that in the general case when no restriction is imposed on the method of scoring, the answer-pattern-differential and score-scatter are still very similar in construction. The scores in such tests are no longer unig, the candidates differing in their estimates of what are the easiest questions, so that a degree of randomness enters into the compiling of scores; this randomness is termed 'hig' in the discussions which follow.

( The terms unig and hig were suggested by Professor Thomson)

## 5. Methods of measuring the relation.

It is one thing to recognize that there is a relation between the answer-pattern-differential and the score-scatter of a test even under hig conditions; it is a much more difficult thing to measure the strength of that relationship. No completely satisfactory method of measuring the relation shown in the results of single tests has yet been devised, though several attempts are reported in a later chapter.(Chap.8.) On the other hand, when the results of many tests given to the

same population are available, it is possible to apply reliable methods of measuring the extent of the relation. The generosity of educationists in Scotland, England, and the United States made it possible for the author to have access to such data.

It will be shown in chapter 5 that the two distributions under consideration have necessarily the same mean. This suggests that the degree of relationship between them may conveniently be measured by correlating certain simple statistics of the two distributions. These used were (i) the standard deviations, (ii) the coefficients of skewness. The calculation of reliable correlation coefficients demands a fairly extensive amount of data, and the collection and examination of suitable data have provided some of the major problems of the work.

6. Experimental results and their implications.

It is found that the correlations between corresponding statistics are positive and fairly high.

(i). The correlation of the standard deviations of the answer-pattern-differential and score-scatter is of the order 0.8 . A discussion of the probable error is deferred to the appropriate chapter. An answer-pattern-differential with a large standard deviation thus tends to produce a score-scatter with a large standard deviation. Such an answer-pattern-differential is derived from an answer-pattern where the items are all of the same degree of difficulty, a type which may be

called flat. On the other hand, a score-scatter with a small standard deviation is the most probable result of using an answer-pattern-differential with a small standard deviation, and this in turn is derived from an answer-pattern some of whose items are very easy, and the remainder very difficult. The intermediate case arises when the answer-pattern used is of the steep type, i.e., one in which the items progress smoothly from very easy to very difficult.

(ii) The correlation of coefficients of skewness of answer-pattern-differential and of score-scatter is found to be of order 0.6 to 0.8 . Skewness of score-scatter is a tool used by examiners to space out candidates either at the upper or lower end of the mark scale. A positively skewed score-scatter spaces out the candidates at the top of the scale. To produce such a skew score-scatter, it is therefore best to use a positively skewed answer-pattern-differential and this is derived from an answer-pattern falling steeply at first, and then flattening out so that the last item is answered by relatively few candidates. Conversely a negatively skewed score scatter may be most easily produced by using an answer-pattern falling gently at first, then steepening.

Thus it is possible to have two tests with identical easy first items and identical most difficult items, yet skewing the score-scatter in opposite directions by reason of the different ways in which the items progress in difficulty. Thus

a test will not necessarily show a positively skewed score-scatter because it has one or two difficult items.

It is also possible to construct a test with a positively skewed answer-pattern-differential, spacing out the top candidates, yet not containing absurdly difficult items. At the same time it must be pointed out that $n_o - n_1$ and $n_m - 0$ are both parts of the answer-pattern-differential and it is difficult to skew this positively if $n_m$ is too large, i.e. if the hardest item is too easy, or negatively if $n_o - n_1$ is too large, i.e. if the easiest item is too difficult. For a given skewness the process of lowering the difficulty of the hardest items, with corresponding adjustments to the other items to keep the skewness constant, reaches a limit in the form of a flat test. In the case of a positively skewed answer-pattern-differential, the items of this flat test are of more than average difficulty, n being less than $\frac{1}{2}n_o$ for all the items; and in the case of a negatively skewed answer-pattern-differentia the items are of the easy type, n being greater than $\frac{1}{2}n_o$ for all the items.

This property of the difficulty of tests in determining skewness of score-scatter has long been known to examiners, and has been used by them to produce score-scatters spacing out the best or the poorest candidates . It will be proved in chapter 7 that the examiner's empirical rule is merely an approximation to the more general principle that skewness of score-scatter is

highly correlated with skewness of answer-pattern-differential It will also be shown that, in general, a more reliable prediction of skewness of score-scatter is obtained from the skewness of the answer-pattern-differential than from the difficulty level alone.

In conclusion, the examiner desiring a particular type of score-scatter must choose his items so that the answer-pattern formed gives an answer-pattern-differential with the required characteristics. If the score-scatter desired is one with a large scatter and a slight positive skewness, then the answer-pattern-differential must also have these characteristics, that is, the answer-pattern must be of the flat type with the items of slightly more than average difficulty. To obtain a score-scatter of average spread, and no skewness, the examiner should use a steep test, the items of which increase uniformly in difficulty from very easy to very difficult. The most convenient way to determine the nature of the answer-pattern required is to use a diagrammatic method.

## 7. The influence of hig.

The exact relationship of answer-pattern-differential and score-scatter vanishes when hig enters and is replaced by a degree of correspondence, such as has been measured by the correlation coefficients of the last section. We should expect the degree of correspondence to be high when the amount of hig

present is small, and to diminish as the incidence of hig becomes greater until with maximum hig there is no relation. This was the view put forward in the author's first paper on the subject. More recent work using more extensive data has created doubt as to whether these conclusions are valid in the form then stated.

We have as yet no absolutely satisfactory method of measuring hig, but it is easy to prove that hig is more likely to occur in the flat type of test, and that a greater probability of unig is obtained from the steep type of test. A coefficient has been devised which ranges in order of steepness tests containing the same number of items. It is shown in chapter 9 that when the tests used are divided by this means into two groups of steeper and flatter tests respectively, the degree of correspondence between answer-pattern-differential and score-scatter is as great with the flatter tests as with the steeper tests. It seems that once the bond of unig has gone only a slight degree of hig is necessary for the answer-pattern-differential and score-scatter to settle into a degree of correspondence which is the same for all tests independent of the incidence of hig.

More probably the explanation is that we are working all the time with a much greater degree of hig than we suspected It is noteworthy that the probability of unig depends not only on the steepness of the test in the sense of making full use

of the range of difficulty, but also on the number of items, diminishing rapidly as the number is increased. For tests with more than two or three items, the probability of hig is great, even for the steepest answer-pattern possible with that number of items. That probability, already high, is only slightly increased by changing the answer-pattern from steep to flat.

The effect of varying degrees of hig on the correspondence between answer-pattern-differential and score-scatter is therefore likely to be confined to tests with very few items. The deductions on the relation of hig and strength of correspondence made earlier and mentioned at the beginning of this section are probably correct only when applied to tests of this kind. In practically every test with which an examiner is likely to deal, the degree of correspondence between answer-pattern-differential and score-scatter is independent of the steepness of the paper.

Apart altogether from the question of steepness and from the limitations imposed by considering each item as independent, it is still possible for an examiner to reduce the amount of hig in a paper by linking items, so that a correct answer to one is only possible when the previous one is correctly answered; by suitable arrangements of multiple marks; and by the format of the paper including directions such as "Begin at the beginning and go straight through".

## 8. Application to school examinations.

It may be appropriate to mention here that the above theory is probably applicable to school examinations of the more usual type where the paper consists of five or six questions, each carrying a number of marks. The small amount of preliminary investigation that has been done by the author shows that the situation is very complicated. Difficulties in the construction of answer-patterns arise from the practices of giving choice of questions, and of awarding marks not only for the contents of the answer but also for the style. At the present stage it is impossible to do more than indicate the probability that here also, answer-pattern suitably defined has an influence on the score-scatter.

## 9. Introduction to subsequent chapters.

The preceding account of theory and results has been necessarily rather brief and compressed, as all the points mentioned are dealt with in much greater detail in the chapters following. In chapter 2, the basic principle is illustrated by examples; thereafter the data used are described ( chapter 3 ); the assumptions underlying the idea of the answer-pattern are considered and their validity tested (chapter 4); and the statistical investigation into the relation of answer-pattern-differential and score-scatter is given in full (chapters 5,6,7); The remaining chapters deal with related points, the last being concerned with the relation of the answer-pattern to the reliability coefficient of the test.

## Chapter 2. The Relation of Answer-pattern and Score-scatter in a Special Case ( the Unig Case ).

### 1. Two points of exact relationship.

That there are some points of exact relationship between the answer-pattern and the score-scatter is obvious. For example, the total number of candidates, $n_o$, is equal to

$$N_o + N_1 + N_2 + \ldots\ldots + N_m,$$

where $N_x$ is the number of candidates scoring exactly x points, and m is the total number of scoring points.

Again, the total number of marks scored by all equals

$$0.N_o + 1.N_1 + 2.N_2 + \ldots\ldots\ldots + m.N_m.$$

and also equals

$$n_1 + n_2 + n_3 + \ldots\ldots\ldots\ldots + n_m.$$

Using the usual summation formulae we may write these identities as

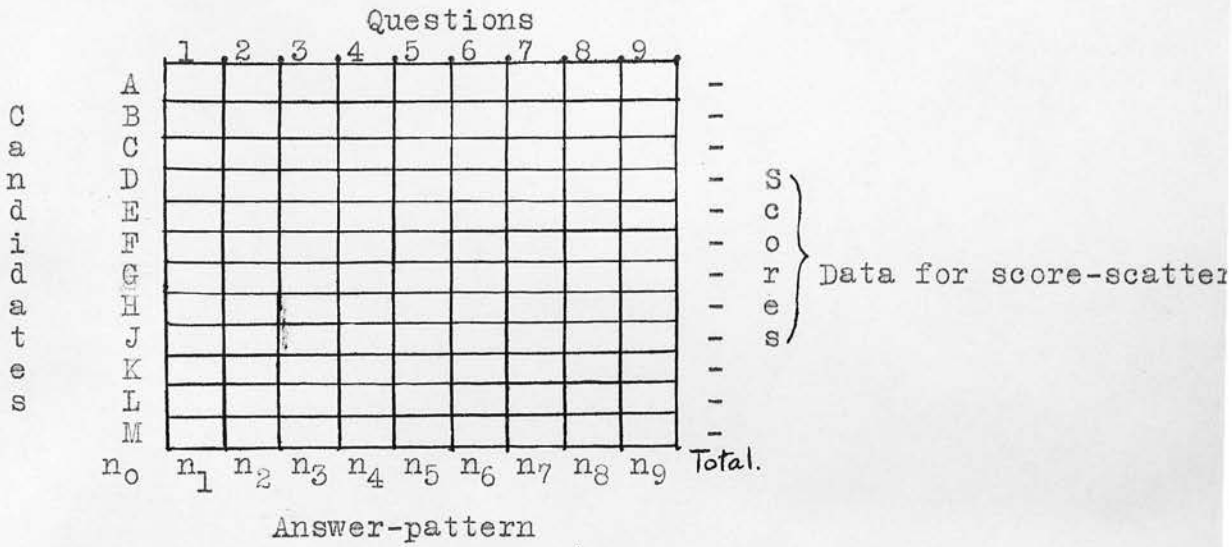
$$\sum_{x=0}^{m} N_x = n_o$$

$$\sum_{x=0}^{m} xN_x = \sum_{x=1}^{m} n_x$$

The second equation provides a useful check in some of the subsequent calculations.

We may tabulate the results of any test in the form shown below, indicating by a mark in the appropriate square a correct answer to a question and from the table read off the answer-pattern and the data for the score-scatter. The sum at the

right hand bottom corner embodies the check mentioned above; the sum of the vertical column above it ($\Sigma xN_x$) must equal the sum of the horizontal row to the left ($\Sigma n_x$).




2. Hig and Unig.

A score of exactly x in a test such as the above may be made up in many ways. Theoretically the most probable composition is that obtained by answering the x easiest question but owing to individual differences, which are smoothed out in the process of adding the columns to get the n's, not every score x is composed in this unique manner. An element of randomness or "higgledy-piggledyness" may enter into the composition of every score except the maximum score, which includes all the available items and must therefore be uniquely composed.

Professor Godfrey Thomson has suggested for this element of higgledy-piggledyness the name "hig", and for the converse of hig the term "unig". These terms are used throughout this paper

At this point it would be as well to state that unig must be taken as meaning more than that every score is made up in a unique manner. It is easy to demonstrate that such a definition would imply linkage between items, a factor we have ruled out. A unig score of x is one then which is composed of answers to the x easiest questions.

3. Relation of answer-pattern (n) and score-scatter (N) under unig.

If hig be excluded, a definite and exact numerical relation can be set up between n and N. Every candidate who scores x exactly must have answered correctly the x easiest questions and no others. The items are arranged so that

$n_0 \geq n_1 \geq n_2 \geq \ldots\ldots\ldots\ldots\ldots \geq n_m$.

In the notation already used,

$N_x$ candidates have answered questions 1,2,3,4,.....,x

$N_{x-1}$ " " " " 1,2,3,4,.....,x-1

and so on, down to

$N_3$ " " " " 1,2,3.

$N_2$ " " " " 1,2.

$N_1$ " " " " 1.

The set of equations may also be continued up to the top equation, which states that

$N_m$ candidates have answered questions 1,2,3,4,.....,m.

By simple addition it is seen that $n_1$, the number of times

question 1 has been correctly answered is equal to

$$N_1 + N_2 + \ldots\ldots\ldots\ldots + N_m.$$

So we derive the set of equations

$$\left.\begin{aligned}
n_o &= N_o + N_1 + N_2 + \ldots\ldots\ldots\ldots + N_m. \quad (1)\\
n_1 &= \quad N_1 + N_2 + \ldots\ldots\ldots\ldots + N_m. \quad (2)\\
n_2 &= \quad\quad N_2 + \ldots\ldots\ldots\ldots + N_m. \quad (3)\\
&\cdot\\
&\cdot\\
n_x &= \quad\quad\quad N_x + N_{x+1} \ldots + N_m. \quad (x+1)\\
&\cdot\\
&\cdot\\
n_m &= \quad\quad\quad\quad\quad\quad N_m. \quad (m+1)
\end{aligned}\right\} \text{A}$$

An alternative form of these equations may be obtained by subtracting equation 2 from equation 1, equation 3 from equation 2, and so on. This gives the set of equations

$$\left.\begin{aligned}
n_o - n_1 &= N_o\\
n_1 - n_2 &= N_1\\
n_2 - n_3 &= N_2\\
&\ldots\ldots\ldots\\
n_{m-1} - n_m &= N_{m-1}\\
n_m &= N_m
\end{aligned}\right\} \text{B}$$

An alternative method of deriving the equations may make the position clearer. The candidates scoring x have answered correctly questions 1 to x. They fail at question x+1. which

will therefore be answered correctly by those scoring x+1 or more. That is, the number of candidates ($N_x$) who scored x is the number who completed their score at x, and failed to go further, i.e., it is $n_x - n_{x+1}$ . This is the series of equations B. Equations A can be derived by addition of the requisite equations of B.

These equations show that when hig is absent, a given answer-pattern completely fixes the score-scatter. This may be illustrated by diagrams, as is done below.

These diagrams show the answer-pattern and score-scatter for a test of 8 questions sat by 22 candidates. Strictly, the answer-pattern is not a curve, but a series of discrete points with values only for the integral values of the abscissae; for the sake of clarity these have been joined. The score-scatter is shown in the familiar histogram form.

In the diagram the equations

$$n_o - n_1 = N_o \ , \quad n_1 - n_2 = N_1 \text{ etc}$$

are shown by AB = A'B', CD = C'D' .....

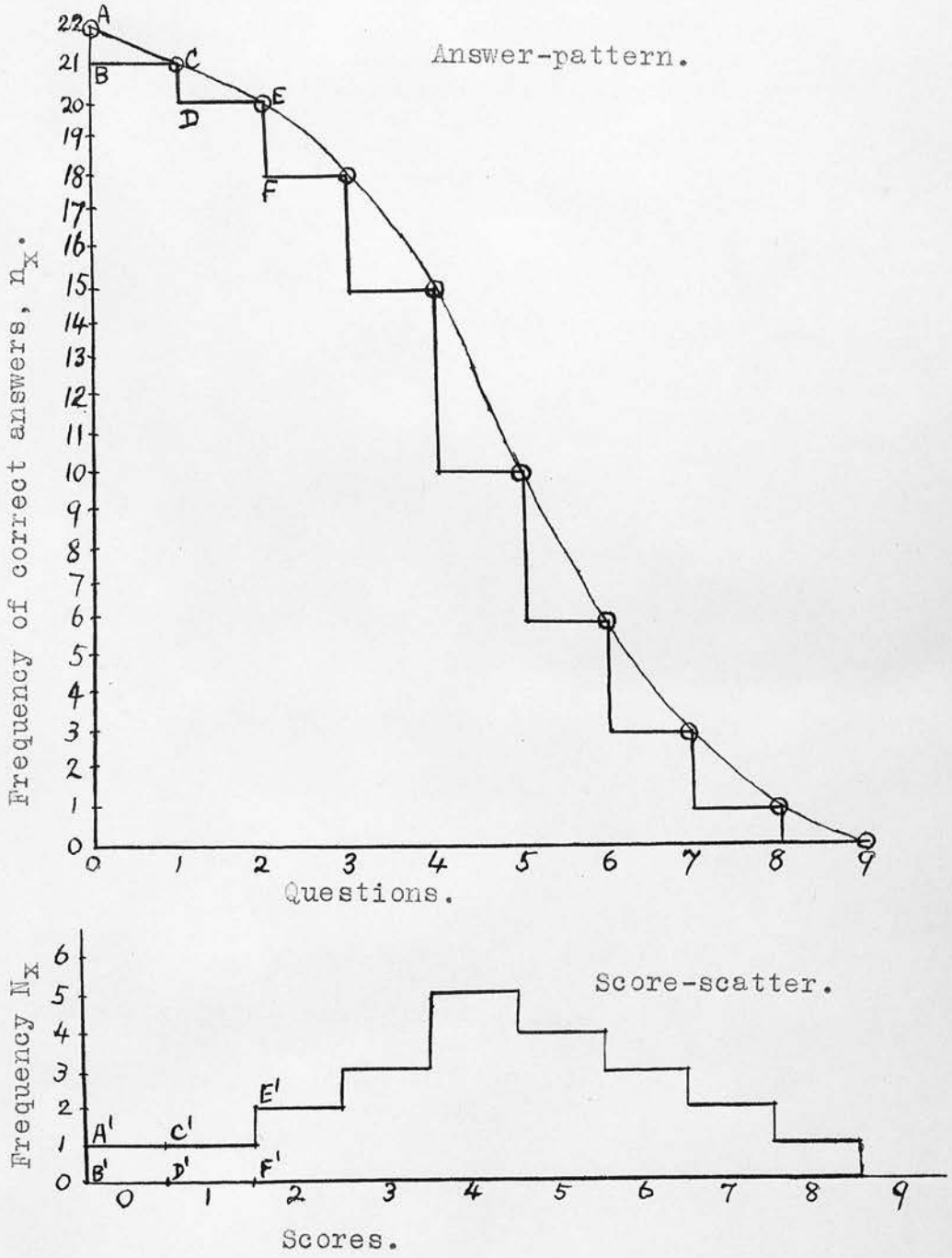



Figure 1. Diagrams showing the relation between answer-pattern and score-scatter in the unig case.

To illustrate the effects of different types of answer-pattern on the score-scatter, the following four fictitious sets of results of tests were constructed. These tests are all 12-item tests, attempted by 100 candidates. The methods used to construct the score-scatters and answer-patterns are explained in an appendix to this chapter, as an elaboration of the mathematical methods employed might obscure the main line of reasoning of the chapter.

The answer-patterns and score-scatters of these four tests are given overleaf, and are graphed on the following page, corresponding curves being indicated by the appropriate numerals As before, curves have been used for the sake of clarity.

Curve 1 represents a normal distribution with mean score 6, and standard deviation 2.

Curve 2 represents a normal distribution with the same mean score 6, but a smaller standard deviation, equal to unity.

Curve 3 is a skew curve, with mean score 6, standard deviation 2, and skewness,as measured by the standardised third moment, equal to 0.66 .

Curve 4 is skewed in the opposite direction; its mean is 6, its standard deviation is 2, and its skewness is -0.66 .

Table 1.

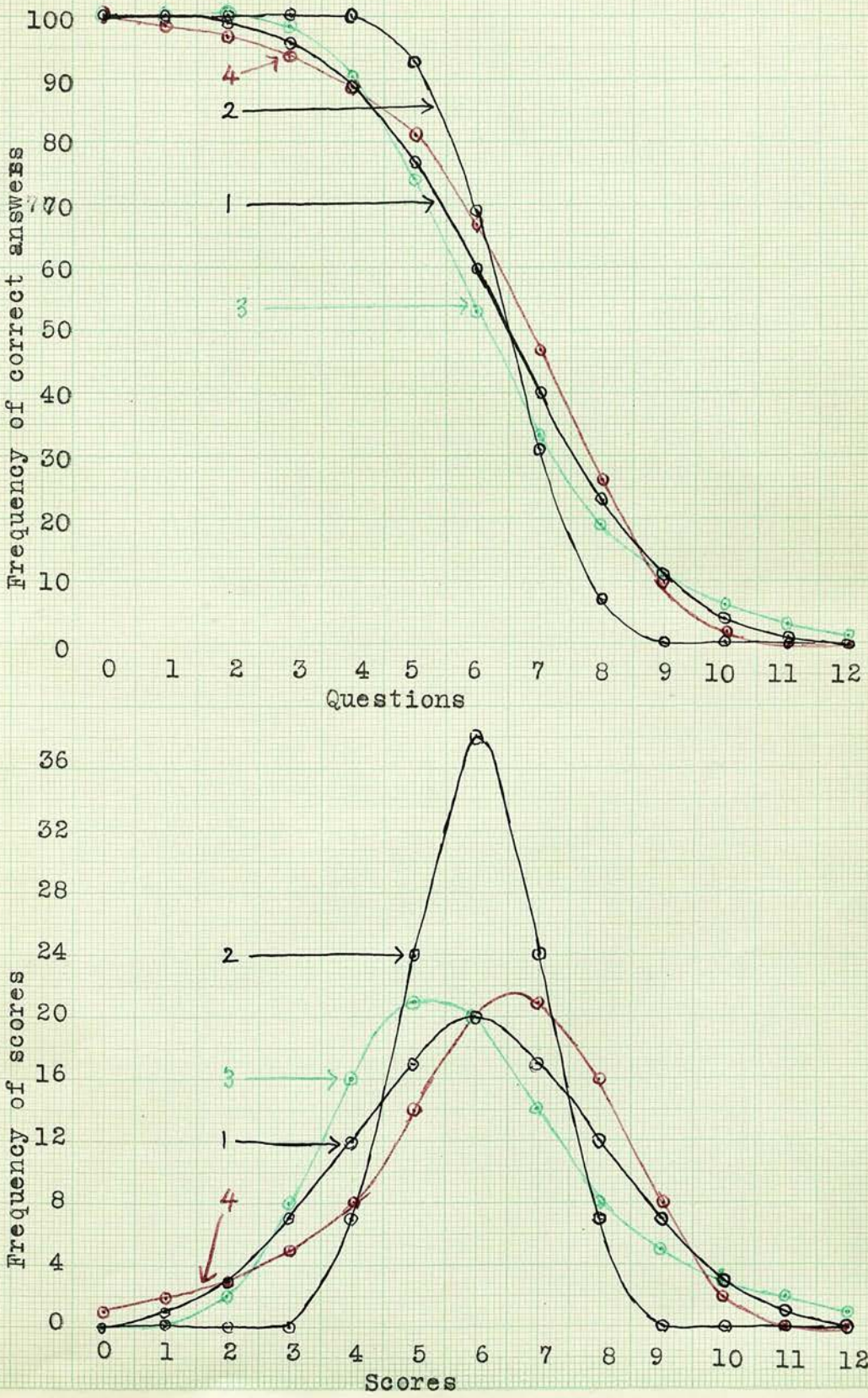
Answer-patterns and score-scatters for tests 1,2,3,4.

$N_x$ = number of candidates scoring x

$n_x$ = number of times item x is correctly answered.

| | Test 1 | | Test 2 | | Test 3 | | Test 4 | |
|---|---|---|---|---|---|---|---|---|
| x | $N_x$ | $n_x$ | $N_x$ | $n_x$ | $N_x$ | $n_x$ | $N_x$ | $n_x$ |
| 0 | 0 | 100 | 0 | 100 | 0 | 100 | 1 | 100 |
| 1 | 1 | 100 | 0 | 100 | 0 | 100 | 2 | 99 |
| 2 | 3 | 99 | 0 | 100 | 2 | 100 | 3 | 97 |
| 3 | 7 | 96 | 0 | 100 | 8 | 98 | 5 | 94 |
| 4 | 12 | 89 | 7 | 100 | 16 | 90 | 8 | 89 |
| 5 | 17 | 77 | 24 | 93 | 21 | 74 | 14 | 81 |
| 6 | 20 | 60 | 38 | 69 | 20 | 53 | 20 | 67 |
| 7 | 17 | 40 | 24 | 31 | 14 | 33 | 21 | 47 |
| 8 | 12 | 23 | 7 | 7 | 8 | 19 | 16 | 26 |
| 9 | 7 | 11 | 0 | 0 | 5 | 11 | 8 | 10 |
| 10 | 3 | 4 | 0 | 0 | 3 | 6 | 2 | 2 |
| 11 | 1 | 1 | 0 | 0 | 2 | 3 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |

Fig. 2. Answer-patterns and Score-scatters of Tests 1,2,3, and 4.



Remembering always that these curves have been obtained in the special case of unig tests, we may note the following very interesting points.

The normal distribution 1 has an answer-pattern which is not frequently met with in tests and examinations. Roughly it may be said to contain four very easy questions, four difficult questions, and only two questions of approximately average difficulty. The normal distributions met with in many score-scatters are not always, or even often, accompanied by this type of answer-pattern, We might proceed to infer that the normal character of these score-scatters occurs in spite of the nature of the answer-pattern; or alternatively, that this shows that there is no strong relation between the answer-pattern and the score-scatter when unig is absent; but it is better to leave the problem at present, and examine the further information at our disposal.

A second point concerns the spacing out of the candidates at the top, shown in test 3 compared with test 1. This spacing is due to the positive skewness of the score-scatter. Its existence may be demonstrated by a comparison of the numbers making the best scores in the two tests. For this purpose it does not matter whether the top score was the maximum possible, or was different in the two tests; all that concerns us is the distribution of the candidates into the groups we might call best, second best, and so on. Let these groups be denoted

1,2,3,4.....Then from the table on page 23 we derive the following.

Table 2.

Numbers of candidates in various groups

| Group | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Test 1 | 1 | 3 | 7 | 12 | 17 | 20 | 17 | 12 | 7 | 3 | 1 |
| Test 3 | 1 | 2 | 3 | 5 | 8 | 14 | 20 | 21 | 16 | 8 | 2 |
| Test 4 | 2 | 8 | 16 | 21 | 20 | 14 | 8 | 5 | 3 | 2 | 1 |

An inspection of these figures shows that while tests 1 and 3 both pick out the best candidate, from that point onward test 3 is superior in spacing out the best candidates. The 11 best candidates are spread over four groups in test 3, as compared with three groups in test 1. Test 4, on the other hand is superior to test 1 in spacing out the poorer candidates from each other.

The interesting thing about the answer-patterns of the tests 1 and 3 is that this spacing has been accomplished by the small gradation in difficulty from an item too difficult to be done by any candidate through the three or four most difficult items. There is no great difference in difficulty between the hardest question of test 1, and the hardest question of test 3. It is the progression from question to question that apportions the candidates to their respective scores. At the same time it must be noted that this progression includes the step from $n_m$ to

zero; in both 1 and 3 this step is small, because the hardest question is very difficult.

If the number of items in a test be considered to increase indefinitely, then the score-scatter tends to a limit as the differential of the answer-pattern curve. This is expressed by the set of equations B, which may be written in the limit when a great number of questions is considered as

$$-\frac{dn}{dx} = N.$$

That is, the slope of the answer-pattern is minus one times the ordinate of the score-scatter at the corresponding point. $\frac{dn}{dx}$ is the answer-pattern-differential. and this is a useful term to employ for the distribution

$$n_o - n_1,\ n_1 - n_2,\ \ldots\ldots\ldots\ n_m - 0,$$

in the case where m is finite.

The converse of this statement is that the answer-pattern is the integral of the score-scatter. This may be proved by integrating the above equation or by restating the formula A which in the limit becomes

$$n_x = \int_x^\infty N dx$$

If the curve sketched below is $y = N_x$, then $n_x$ equals the shaded area. This conception of the relation of answer-pattern and score-scatter may be found useful during the later work.

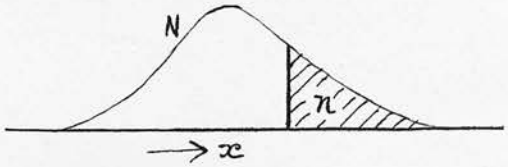



Fig. 3. Answer-pattern as integral of score-scatter.

There are two important particular cases. One is where the answer-pattern is a straight line joining the points $x = 0$, $y = n_0$ and $x = m + 1$, $y = 0$. The score-scatter is then $N_x$ = constant; a rectangular distribution sketched overleaf. This may be readily deduced from the above discussion.

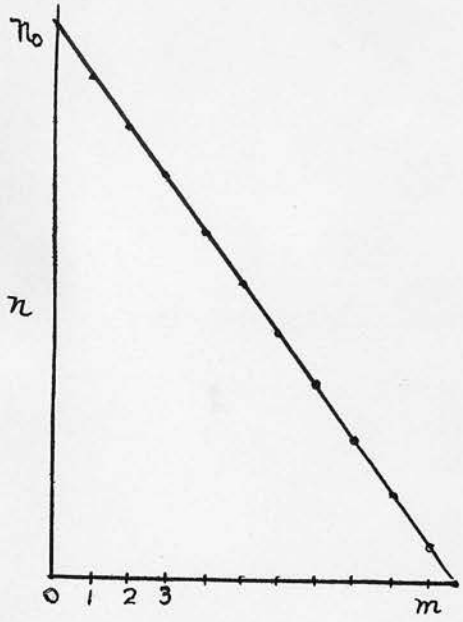
Similarly when the answer-pattern is the line $y = n$, except for the isolated point $x = 0$, $y = n_0$, we have a peculiar and rather important score-scatter, given by the two equations

$$N_0 = n_0 - n$$

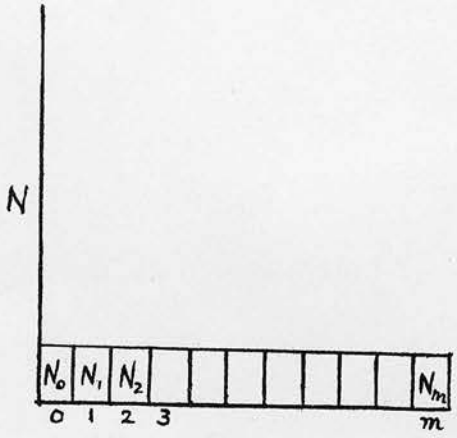
$$N_m = n$$

all the other values of N vanishing, since they are equal to $n - n$. Thus all the candidates score either zero or the maximum, when the test consists of m items all of the same difficulty.

Figure 4. Particular cases discussed on page 28.

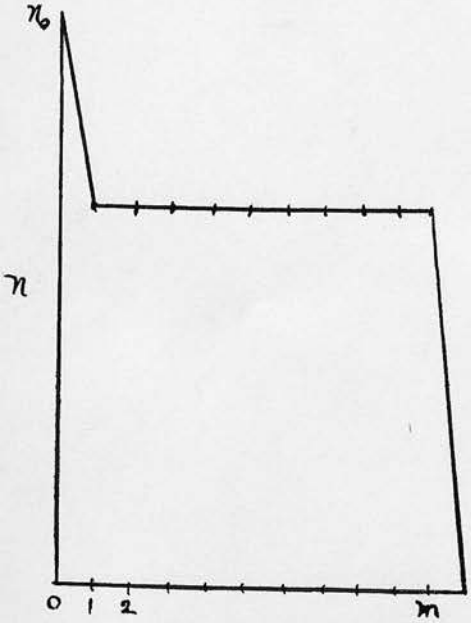



Items.

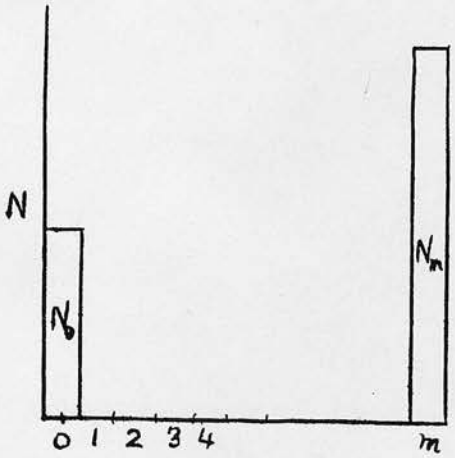



Scores.

Case One




Items.




Scores.

Case Two

4. The occurrence of hig.

Unfortunately for the exactness of the above theory, hig enters in varying degrees into the composition of every test result. Owing to individual variations in preference or ability for certain questions, all scores of x are not made up uniquely of answers to the x easiest questions. These individual variations are smoothed out in the process of adding the columns giving the answer-pattern, but they serve to destroy, in greater or less degree, the exactness of the relationship between answer-pattern and score-scatter, according as that amount of hig present is large or small.

The amount of hig present depends on various factors. For instance if each question in a test were so linked with the preceding question that a correct answer could only be given if the preceding questions had been answered correctly; then the amount of hig present would be zero. The same result might be achieved if the steps of difficulty between the questions were so large that a correct answer to a difficult question would almost certainly be accompanied by a correct answer to the much easier questions. On the other hand, a test whose questions are all of the same difficulty will, it is likely, tend to show maximum hig. Lastly the injunction "Begin at the beginning and go straight through" usually found in intelligence tests will tend to decrease hig.

The first method of decreasing hig is out of court in

this discussion. The last has psychological effects which we cannot attempt to measure here, though it will be mentioned again in connection with intelligence tests. The second factor merits a more detailed investigation.

Since large steps of difficulty between questions will be shown in a steep answer-pattern graph, we may design a test with such large steps as 'steep' and the opposite type of test as 'flat'. A perfectly flat test will have items all of the same degree of difficulty. This is far from being an exact definition of steepness, but the whole question of steepness will be treated more fully later. We shall now prove that hig is more probable with a flat test and is less probable with a steep test.

5. The relation of hig and steepness.

Consider a test made up of m questions, denoted a,b,c,.. not necessarily in order of difficulty. Let the probability of answering question a be $p_a$, and so on. Then $p_a = n_a/n_o$.

The probability that a score of exactly x be made on the test, by answering correctly questions a,b,c,.....k
x in number, is

$$p_a p_b p_c \ldots\ldots p_k (1-p_{a'})(1-p_{b'}) \ldots\ldots$$

where a',b',..... denote the m-x questions unanswered.

The total number of ways of scoring x exactly is expressed by the number of ways x p's and m-x (1-p)'s can be picked out of

the array of m p's. That is, there are mCx ways of scoring exactly x.

Suppose that the questions arranged in ascending order of difficulty are denoted 1,2,3,..... m. Then the probability of scoring exactly x in the unig way, i.e. by answering the x easiest questions is

$$p_1 p_2 p_3 \cdots\cdots p_x (1-p_{x+1})(1-p_{x+2}) \cdots\cdots (1-p_m)$$

Therefore by the use of Bayes' Theorem, we find that the probability that a score x is unig is

$$\frac{p_1 p_2 p_3 \cdots\cdots p_x (1-p_{x+1})(1-p_{x+2}) \cdots\cdots (1-p_m)}{\sum p_a p_b p_c \cdots\cdots p_k (1-p_{a'})(1-p_{b'}) \cdots\cdots}$$

Substituting $p_x = n_x/n_o$ and eliminating the $n_o$ denominators we obtain for the probability of unig in the score x

$$u_x = \frac{n_1 n_2 n_3 \cdots n_x (n_o - n_{x+1}) \cdots\cdots (n_o - n_m)}{\sum n_a n_b n_c \cdots n_k (n_o - n_{a'}) \cdots\cdots (n_o - n_{k'})}$$

The probability that a whole test is answered in unig fashion is the probability that every score in the test is unig. This probability equals $\prod_1^m u_x$, and the probability of hig is $1 - \prod u_x$.

The n's form a descending monotone sequence of positive numbers. It is therefore clear that the probability increases rapidly as m is increased. For a given value of m, the maximum and minimum values of the function will obviously occur (1) when all the n's (not including $n_o$) are equal to each other; and

(2) when the n's are as far as possible from being equal to each other. Which of these states corresponds to the maximum and which to the minimum must now be ascertained.

When $n_1 = n_2 = n_3 = \ldots\ldots = n_m = n$ say, $\prod u_x$ becomes

$$\prod_{x=1}^{m} \frac{n^x(n_o-n)^{m-x}}{\sum n^x(n_o-n)^{m-x}} = \prod_{x=1}^{m} \frac{1}{m\ C\ x} = \frac{(1!2!\ldots(m-1)!)^2}{(m!)^{m-1}}$$

This decreases very rapidly as m grows, being as low as 0.11 for m = 3. It corresponds to the minimum value of $\prod$ so that the probability of hig is a maximum for the given m.

For the probability of hig to be small it is necessary that the product $\prod$ should tend to unity. That is,

$\sum n_a n_b n_c \ldots (n_o-n_{a'})(n_o-n_{b'})\ldots\ldots$

tends to

$$n_1 n_2 n_3 \ldots (n_o-n_{x+1})(n_o-n_{x+2})\ldots\ldots(n_o-n_m)$$

What is required then is that the sum of all the other terms except the above in the summation must be small in comparison with the term considered. These other terms may be distinguished from the first by the fact that the total of the subscripts of the n's which are multiplied together is a minimum for the first term. It is then obvious that the sum considered will be a minimum if $n_1$ is much greater than $n_2$ and so on, i.e.

$$n_1 \geqslant n_2 \geqslant n_3 \geqslant \ldots\ldots\ldots\ldots n_{m-1} \geqslant n_m.$$

This corresponds to what we have called a 'steep' test.

On a priori grounds we therefore find the expectation of hig to be greater with a flat test than with a steep test, provided that both tests have the same number of items. When the number of items varies, the test with the larger number of items has already on that account a greater expectation of hig. It would have been interesting to have compared the expectations of hig for two tests, the first having few items but of a flat nature, the second having many items, but these arranged in as steep an answer-pattern as possible. Unfortunately for this, the calculation of the a priori probability of hig is extremely laborious for any save the shortest tests. For a 10-item test it would involve the calculation of over a thousand products each of ten ratios.

## Appendix. The Methods of Construction of Score-scatters used in Tests 1,2,3,4.

### Tests 1 and 2.

The standard form of the normal frequency curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

where a = mean value of x,

$\sigma$ = standard deviation of x from mean.

The frequency of scores between $x_1$ and $x_2$ is then given by the integral

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{(x-a)^2}{2\sigma^2}} dx$$

This distribution is continuous, but the scores required are integers. The frequency of a score 5 must therefore be measured by the frequency between the ordinates at x = 4.5 and x = 5.5 . The points of demarcation (x) to be used are thus midway between the integers.

By the substitution $X = \frac{x-a}{\sigma}$ , the above integral becomes

$$\frac{1}{\sqrt{2\pi}} \int_{X_1}^{X_2} e^{-\frac{X^2}{2}} dx$$

where $X_1$ and $X_2$ are the new limits. The integral may now be written

$$\frac{1}{\sqrt{2\pi}} \int_0^{X_2} e^{-\frac{X^2}{2}} dx \quad - \quad \frac{1}{\sqrt{2\pi}} \int_0^{X_1} e^{-\frac{X^2}{2}} dX .$$

These are examples of the well-known probability integral which has been tabulated. The tables used in the calculation below were in Pearson's Tables for Statisticians and Biometricians.

The steps of the method are then;

(1) evaluate $X = \frac{x-a}{\sigma}$,

(2) find from the tables the values of $I = \frac{1}{\sqrt{2\pi}} \int_0^X e^{-\frac{X^2}{2}} dX$,

(3) by differences find the relative frequencies,

(4) by multiplying by the number of candidates find the actual frequencies.

In test 1, $a = 6$, $\sigma = 2$, $n_o = 100$. Thus we obtain the table

Table 3

| Score | x | X | I | Differences | Frequencies |
|---|---|---|---|---|---|
| | | $(-\infty)$ | -0.5000 | | |
| 0 | | | | 0.0030 | 0 |
| | 0.5 | -2.75 | -0.4970 | | |
| 1 | | | | 0.0092 | 1 |
| | 1.5 | -2.25 | -0.4878 | | |
| 2 | | | | 0.0279 | 3 |
| | 2.5 | -1.75 | -0.4599 | | |
| 3 | | | | 0.0655 | 7 |
| | 3.5 | -1.25 | -0.3944 | | |
| 4 | | | | 0.1210 | 12 |
| | 4.5 | -0.75 | -0.2734 | | |
| 5 | | | | 0.1747 | 17 |
| | 5.5 | -0.25 | -0.0987 | | |
| 6 | | | | 0.1974 | 20 |
| | 6.5 | +0.25 | +0.0987 | | |
| 7 | | | | 0.1747 | 17 |
| | 7.5 | +0.75 | +0.2734 | | |
| 8 | | | | 0.1210 | 12 |
| | 8.5 | +1.25 | +0.3944 | | |
| 9 | | | | 0.0655 | 7 |
| | 9.5 | +1.75 | +0.4599 | | |
| 10 | | | | 0.0279 | 3 |
| | 10.5 | +2.25 | +0.4878 | | |
| 11 | | | | 0.0092 | 1 |
| | 11.5 | +2.75 | +0.4970 | | |
| 12 | | | | 0.0030 | 0 |
| | | $(+\infty)$ | +0.5000 | | |
| | | | | | 100 |

Inexactly the same way, but using $\sigma = 1$, the score-scatter used in test 2 may be evaluated.

Tests 3 and 4.

Here skew curves are desired. The most suitable statistical aggregate to use is Charlier's type A distribution

$$y = \psi(x) + A_3 \psi'''(x) + A_4 \psi^{IV}(x) + \ldots\ldots$$

where $\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

and the A's are functions of the moments about the mean.

For our purpose a sufficient approximation will be obtained by using two terms of the series. By integrating the above expression for y between the limits $\pm\infty$, it is easy to show that $A_3 = -\frac{\mu_3}{6}$ where $\mu_3 = \int_{-\infty}^{\infty} yx^3\,dx$.

$\mu_3$ is a measure of the skewness of the distribution.

As in the case of the normal curves 1 and 2 we must evaluate the integral $\int_0^x y\,dx$ which equals

$$\int_0^x \psi(x)\,dx - \frac{\mu_3}{6}\int_0^x \psi'''(x)\,dx.$$

The first of this pair of integrals is the probability integral already used. The second integral $\int_0^x \psi'''(x)\,dx$ may be shown to equal $\int_0^x (3x - x^3)\psi(x)\,dx$

which equals $3\int_0^x x\psi(x)\,dx - \int_0^x x^3\psi(x)\,dx.$

Tables of these last two integrals are to be found in Pearson's 'Tables' page 22. What is tabulated is

$$m_1 = \int_0^x x\psi(x)\,dx \quad \text{and} \quad m_3 = \tfrac{1}{2}\int_0^x x^3\psi(x)\,dx$$

so that the expression for $\int_0^x y\,dx$ becomes

$$I = \frac{\mu_3 m_1}{2} + \frac{\mu_3 m_3}{3}.$$

In curve 3, put mean a = 6, $\sigma = 2$, $\mu_3 = +1$.

The method then follows that used with tests 1 and 2, save that two additional terms $-\frac{m_1}{2} + \frac{m_3}{3}$ have to be added. Thus the following table is obtained.

Table 4.

| Score | x | X | I | $\frac{1}{2}m_1$ | $\frac{1}{3}m_3$ | $\int_0^x y\,dx$ | Diffce | Freq. |
|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | |
| | 0.5 | -2.75 | -0.4970 | +0.195 | +0.118 | -0.574 | | |
| 1 | | | | | | | | 0 |
| | 1.5 | -2.25 | -0.4878 | +0.184 | +0.095 | -0.577 | | |
| 2 | | | | | | | 0.021 | 2 |
| | 2.5 | -1.75 | -0.4599 | +0.156 | +0.060 | -0.556 | | |
| 3 | | | | | | | 0.079 | 8 |
| | 3.5 | -1.25 | -0.3944 | +0.108 | +0.025 | -0.477 | | |
| 4 | | | | | | | 0.159 | 16 |
| | 4.5 | -0.75 | -0.2734 | +0.049 | +0.004 | -0.318 | | |
| 5 | | | | | | | 0.213 | 21 |
| | 5.5 | -0.25 | -0.0987 | +0.006 | +0.000 | -0.105 | | |
| 6 | | | | | | | 0.198 | 20 |
| | 6.5 | +0.25 | +0.0987 | +0.006 | +0.000 | +0.093 | | |
| 7 | | | | | | | 0.135 | 14 |
| | 7.5 | +0.75 | +0.2734 | +0.049 | +0.004 | +0.228 | | |
| 8 | | | | | | | 0.083 | 8 |
| | 8.5 | +1.25 | +0.3944 | +0.108 | +0.025 | +0.311 | | |
| 9 | | | | | | | 0.053 | 5 |
| | 9.5 | +1.75 | +0.4599 | +0.156 | +0.060 | +0.364 | | |
| 10 | | | | | | | 0.035 | 3 |
| | 10.5 | +2.25 | +0.4878 | +0.184 | +0.095 | +0.399 | | |
| 11 | | | | | | | 0.021 | 2 |
| | 11.5 | +2.75 | +0.4970 | +0.195 | +0.118 | +0.420 | | |
| 12 | | | | | | | 0.014 | 1 |
| | | | +0.5000 | +0.199 | +0.133 | +0.434 | | |
| | | | | | | | | 100 |

If the skewness of this distribution is calculated in the usual way, it is found to be +0.66 . This discrepancy is due partly to the approximation made by using only two terms of the Charlier series becoming less accurate for values of $\mu_3$ as large as 1. It introduces no complication into the result.

The data for test 4 are obtained by reversing the signs before $\frac{1}{2}m_1$ and $\frac{1}{3}m_3$ before summation.

## Chapter 3. A Description of the Data used.

As has already been indicated, the type of test under discussion at present is one made up of independent items, each carrying unit score when correctly answered. Data of this sort, or even the answer-patterns and score-scatters of such tests, are not easy to obtain in any quantity, and one of the chief difficulties of this investigation has been the obtaining of suitable data. The material used in this paper has been obtained mainly from three sources.

(1) The results of the Moray House Series of Group Tests of Intelligence which have been supplied to certain English boroughs. For the necessary data which have been compiled from these results the author is indebted to Professor Godfrey H. Thomson, and in two cases to the Education Committee for access to the papers themselves.

The tests used were Moray House Tests 8,9,11, and 12. The last consisted of two parts, a verbal and a pictorial, which will be referred to as 12v and 12p. In the subsequent discussion these tests will be referred to as M.H.T. 8,9,11,12v and 12p. The numbers sitting the tests were usefully large.

(2) As part of the course for the degree of Bachelor of Education, the author commenced an investigation into the present problem, and in the course of the experiments constructed three tests, titled A,B,C, which were given to children in two English schools. These tests will be referred to as A,B,C.

(3) A study of Thorndike's "Measurement of Intelligence" suggested that Professor Thorndike might have some data of the type required. Through the good offices of Professor Godfrey Thomson and the courtesy of Professor Thorndike, the author was able to obtain from Columbia University, New York, two extremely useful sets of data.

The first was the complete score sheet of a test of 410 items which had been attempted by 32 candidates who went through a given number each day. This test was split up into 41 sub-tests each of 10 consecutive items. Its great usefulness lies in the fact that here we have a fixed population; the changes in score-scatter may then be studied in relation th the changes in answer-pattern without any interference from population effects. These tests will be referred to as the 41 tests.

The second set consisted of score sheets of 15 tests sat by candidates who differed from test to test. In all cases the answer-pattern and the score-scatter could be found from the score sheet. Ten of these were arithmetic tests, denoted as A II, A IX, etc., while five were word knowledge tests and are denoted K,L,$K_2$,M,$M_2$, following the notation used on the score sheets.

A summary of the data is shown in the table overleaf.

| Test | Number of Items | Number of Candidates |
|---|---|---|
| M.H.T. 8. | 109 | 528 |
| M.H.T. 9. | 94 | 202 |
| M.H.T. 11. | 99 | 209 |
| M.H.T. 12v. | 75 | 1000 |
| M.H.T. 12p | 9 | 900 |
| Thesis A | 15 | 166 |
| Thesis B | 15 | 166 |
| Thesis C | 15 | 166 |
| 41 tests ((each)) | 10 | 32 |
| Complete tests | | |
| A II | 10 | 228 |
| A IX | 10 | 170 |
| A X | 10 | 219 |
| A XI | 10 | 204 |
| A XIII | 10 | 141 |
| A XV | 10 | 224 |
| A XXVI | 10 | 89 |
| A XXXIV | 10 | 77 |
| A XXXII | 10 | 54 |
| A XXXIII | 10 | 55 |
| A | 99 | 124 |
| L | 100 | 122 |
| $L_2$ | 99 | 89 |
| M | 100 | 63 |
| $M_2$ | 99 | 87 |
| | | 5415 |

## Chapter 4. The Permanence of Answer-patterns.

So far we have assumed that the answer-pattern of a test is a permanent characteristic, depending only on the items given and the population tested. Such a hypothesis is quite plausible, but should be tested, if at all possible, by experiment. If the same items were attempted again by the same candidates under the same conditions, would the answer-pattern be substantially the same ? Another problem that must be investigated is the permanence of the answer-patterns of sub-tests obtained by selection of items from the whole test. Can the answer-patterns of these sub-tests be reliably predicted from the pattern of the test from which they are selected ?

There are obvious difficulties in the testing of the permanence of answer-patterns. It is impossible to repeat the test under exactly the same conditions; the very fact that the candidates have already attempted the items makes a great difference. It is possible however to devise ways of testing this hypothesis of permanence. For example, using the same test, we may compare the answer-patterns of two groups of testees as similar as possible in composition. Using the method employed in Pearson's test of Goodness of Fit, we may compare the frequencies of correct answers to all the items in one set of data with the corresponding frequencies in the second set of data, and determine whether any differences are attributable to errors of sampling.

There is one important point to be settled first. The

answer-pattern has been defined as the series of the n's, when the items have been placed in ascending order of merit. Suppose that in the first set of results the items in order of difficulty are numbered 1,2,3,4,.... and the frequencies of the correct answers are $n_1, n_2, n_3, n_4, \ldots$ . If there is no change in that order of difficulty with the second set of results the comparison will be quite straightforward. If however there are changes, e.g. suppose the order of difficulty for the second set is 1,3,4,2,... with corresponding frequencies $n_1', n_3', n_4', n_2', \ldots$, then a decision will have to be made whether to compare frequencies for the same item, i.e. $n_1$ with $n_1'$, $n_2$ with $n_2'$, $n_3$ with $n_3'$,..... or to compare frequencies of corresponding items in the rearranged order, i.e. $n_1$ with $n_1'$, $n_2$ with $n_3'$, $n_3$ with $n_4'$, n4 with $n_2'$,.... It seems that the second method may be justified on the following counts:

(1) It is when the items have been arranged in order of difficulty that we obtain the answer-pattern which, in the unig state, controls the score-scatter.

(2) The fact that comparisons have to be made of different populations probably decreases the goodness of fit from that which would be obtained were it possible to use the same population. This effect will be counteracted to some extent by the use of the second method, which will increase the fit .

## 2. Method of Comparison of Answer-patterns, using Pearson's Coefficient of Goodness of Fit.

The details of the method employed in testing goodness of fit are as follows: Let one answer-pattern ( preferably that obtained from the larger group of candidates ) be regarded as the theoretical or standard distribution, giving a set of values $n_x$. The other answer-pattern, regarded as the observed distribution, provides a second set of frequencies $n_x'$. Then $\chi^2 = \sum \frac{(n_x - n_x')^2}{n_x}$ may be calculated, the summation extending over all the classes. Now the probability that a given value of $\chi^2$ should be exceeded through the effects of random sampling has been tabulated. The tables used in the present calculations are those incorporated in Fisher's "Statistical Methods"- Table 3. The value of $n$ to be used in this table is the number of degrees of freedom in which the observed series may differ from the hypothetical; in other words, it equals the number of classes in which the frequency may be filled in arbitrarily. In the present calculations it equals the number of classes since $\sum n_x$ is not necessarily equal to $\sum n_x'$.

Thus we may find whether any differences found, and measured by $\chi^2$, are attributable to the effects of random sampling. It may be as well to note here that any classes with frequencies less than 5 should be lumped together.

Example 1. As a first example, consider the following results obtained with the test M.H.T. 12p, a 9 item test. The answer-patterns to be compared were obtained from 450 boys and 450 girls, the boys' answer-pattern being taken as standard. The results as obtained, and without any rearrangement in order of difficulty, were as follows.

Table 5. Frequencies of correct answers in M.H.T. 12p.

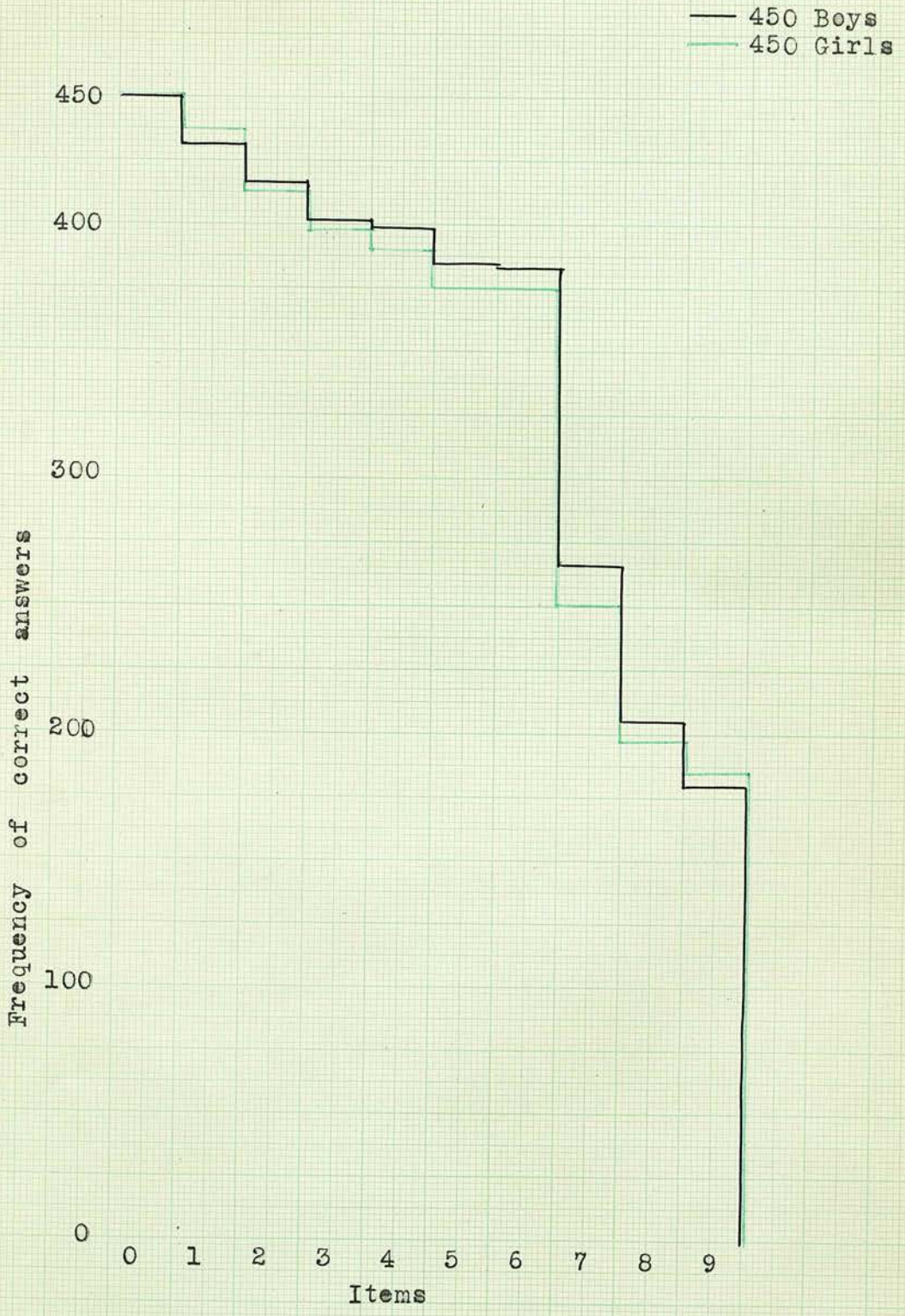
| Item | Boys (n) | Girls (n') | $\frac{(n-n')^2}{n}$ |
|---|---|---|---|
| 1 | 431 | 437 | .09 |
| 2 | 416 | 413 | .02 |
| 3 | 399 | 392 | .12 |
| 4 | 386 | 375 | .31 |
| 5 | 402 | 398 | .04 |
| 6 | 389 | 375 | .50 |
| 7 | 179 | 184 | .14 |
| 8 | 268 | 251 | 1.08 |
| 9 | 205 | 198 | .25 |
| | 3075 | 3023 | $\chi^2 = 2.55$ |

The number of degrees of freedom is 9.

Consulting the table of probabilities, we find that for n = 9, $\chi^2 = 2.55$, the value of P is 0.98 , that is, deviations as large as those found would arise in 98 cases out of 100 in random sampling. Even as they stand, without being arranged as answer-patterns, the two distributions are substantially the same.

The order of difficulty is 1,2,5,3,6,4,8,9,7, in both groups. Rearranging both series in their order of difficulty would therefore make no difference to the above calculation, and the results still hold. The two answer-patterns are substantially the same. They are graphed on the next page.

Fig. 5. Answer-patterns of M.H.T. 12p.



$\sqrt{2\chi^2} - \sqrt{2n - 1} = 22.2 - 12.3 = 9.9$

The deviation equals 9.9 times the standard deviation and is therefore significant. The two sets of frequencies cannot be regarded as substantially the same.

If the items are rearranged in order of difficulty, so as to form the answer-patterns, the value of $\chi^2$ now becomes 79.1 .

$\sqrt{2\chi^2} - \sqrt{2n - 1} = 12.6 - 12.3 = 0.3$ .

The deviation is now only 0.3 times the standard deviation; the differences between the two answer-patterns are well within the limits of error due to random sampling, and the two answer-patterns may be regarded as essentially the same. They are graphed on the following page.

Example 3. The data of M.H.T. 11 provide another application of this test. On page 50 are drawn the answer-patterns of results obtained from two sets of candidates, the first comprising 105 boys and girls, and the second 104 boys and girls from a different school.

The value of $\chi^2$ for these two answer-patterns is 38.0, which is so low a value for a 99 item test that the probability of a larger deviation in random sampling is over 0.99 . The curves are almost identical. It may be that the improvement over the previous examples is due to the elimination of sex differences, each group of candidates containing both boys and girls.

Example 2. The other part of this test, referred to as M.H.T.12v had 76 items. The following data were obtained from the papers of 500 boys and 500 girls sitting the test.

Table 6. Frequencies of correct answers in M.H.T. 12v.

| Items 1-26 | | Items 27-52 | | Items 53-76 | |
|---|---|---|---|---|---|
| Boys | Girls | Boys | Girls | Boys | Girls |
| 467 | 450 | 149 | 143 | 128 | 119 |
| 450 | 456 | 156 | 121 | 276 | 267 |
| 248 | 239 | 373 | 344 | 150 | 138 |
| 456 | 458 | 179 | 172 | 355 | 337 |
| 388 | 401 | 348 | 315 | 236 | 238 |
| 393 | 383 | 376 | 314 | 133 | 130 |
| 357 | 389 | 375 | 347 | 216 | 129 |
| 359 | 338 | 216 | 167 | 137 | 99 |
| 248 | 232 | 119 | 100 | 91 | 81 |
| 280 | 297 | 408 | 408 | 159 | 126 |
| 408 | 408 | 177 | 229 | 162 | 147 |
| 403 | 391 | 193 | 214 | 261 | 255 |
| 396 | 386 | 195 | 179 | 176 | 179 |
| 316 | 315 | 218 | 220 | 190 | 199 |
| 322 | 300 | 389 | 396 | 81 | 87 |
| 382 | 393 | 397 | 391 | 24 | 52 |
| 168 | 172 | 331 | 328 | 34 | 27 |
| 64 | 87 | 188 | 180 | 23 | 24 |
| 308 | 282 | 137 | 124 | 11 | 19 |
| 149 | 141 | 83 | 48 | 232 | 224 |
| 430 | 420 | 278 | 248 | 157 | 137 |
| 197 | 195 | 226 | 163 | 184 | 147 |
| 335 | 312 | 159 | 142 | 53 | 43 |
| 363 | 353 | 175 | 157 | 221 | 220 |
| 348 | 345 | 380 | 366 | 18546 | 17763 |
| 145 | 141 | 251 | 239 | | |

If these frequencies be compared item by item as they stand, the value of $\chi^2$ is found to be 247.4 . The number of degrees of freedom is 76. This is usually denoted by n. For such large values of n, there are no entries in the tables; we use the fact that when n is large $\sqrt{2\chi^2} - \sqrt{2n - 1}$ is normally distributed about zero with unit standard deviation.

Fig 6. Answer-patterns of M.H.T. 12v .

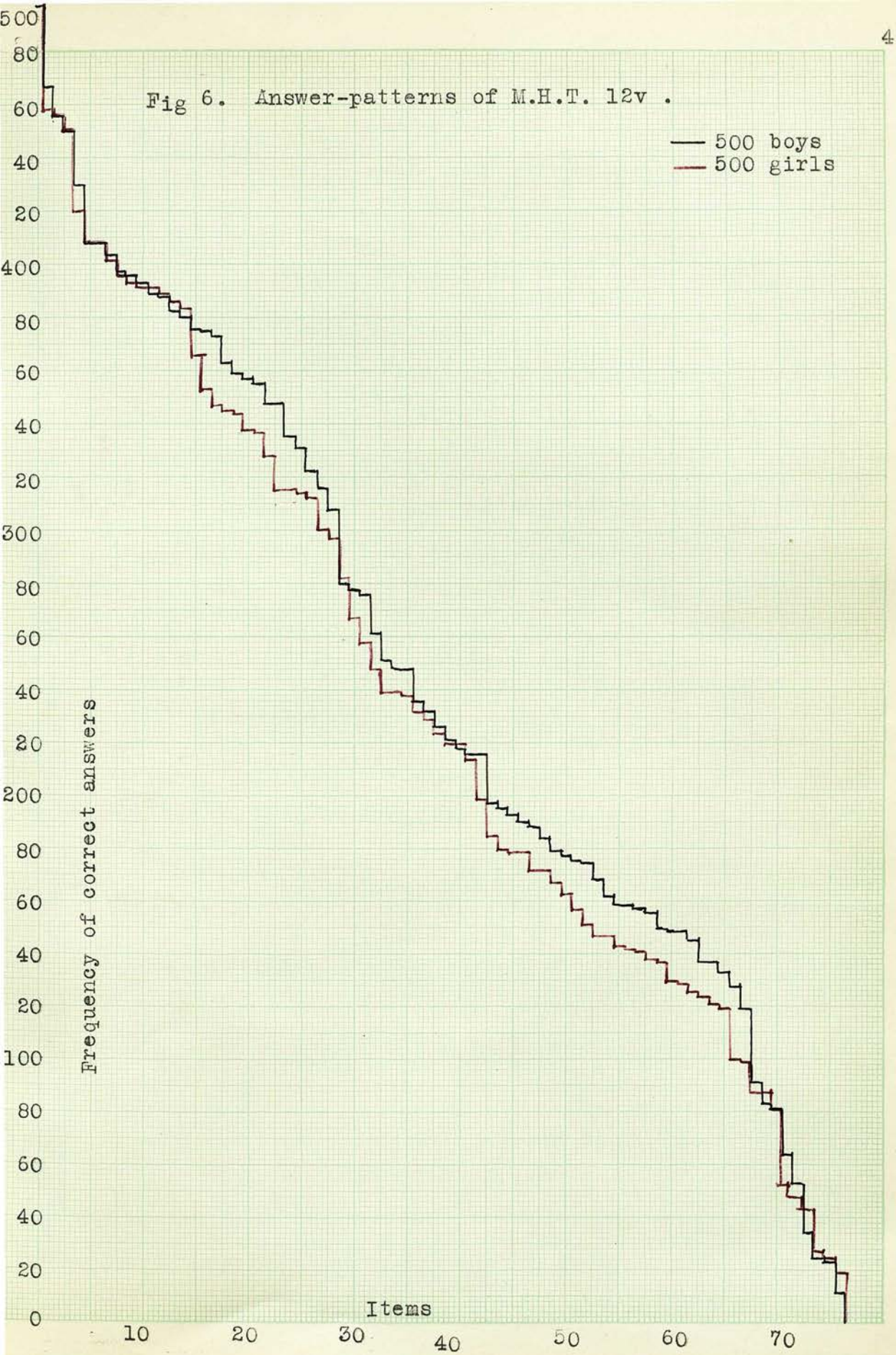
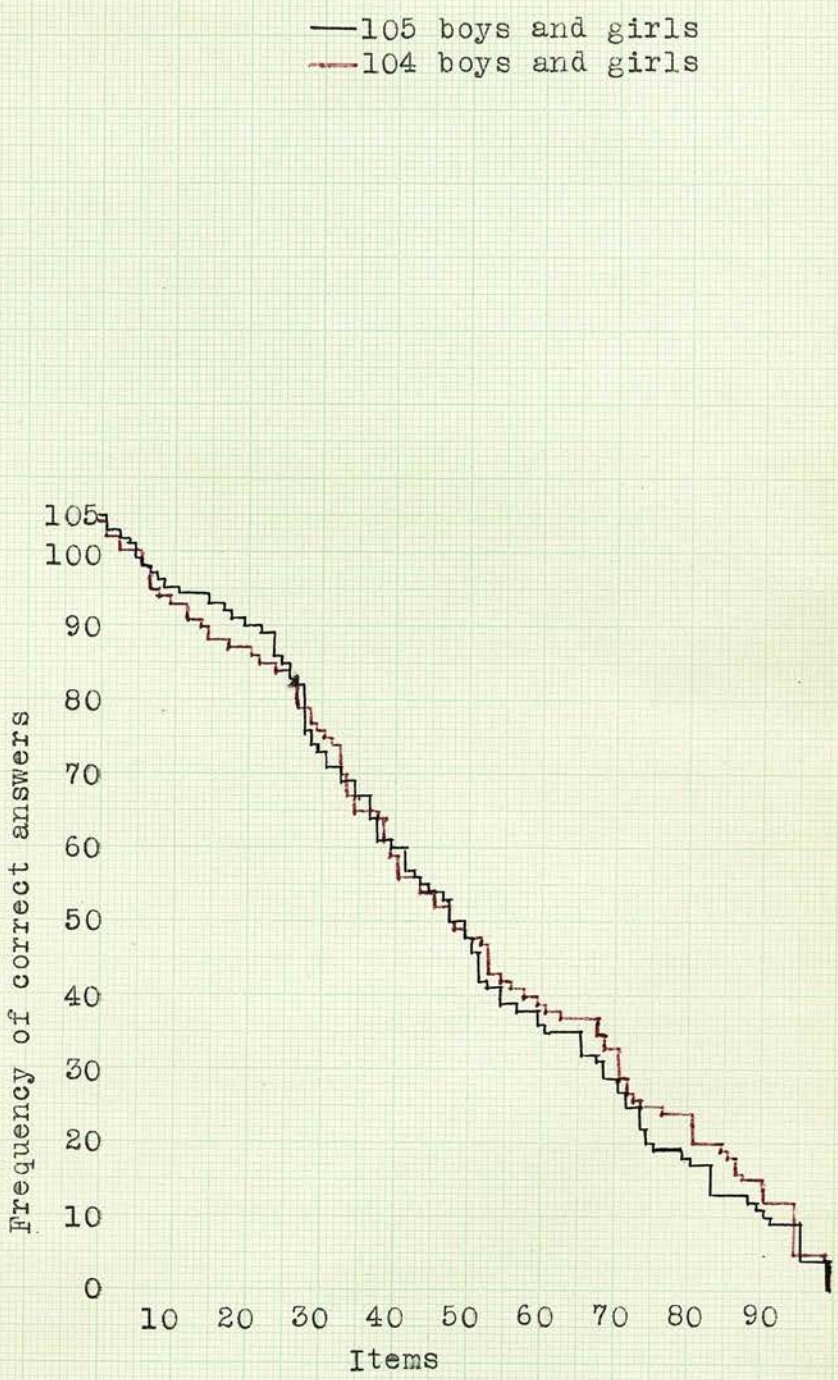


Fig. 7. Answer-patterns of M.H.T. 11.



Example 4. The data of M.H.T. 9 not only provide another example but also raise a very interesting and important point. The calculation is given in full to make this clear. The data were obtained from 105 boys and girls in two schools, giving set A, and 97 boys and girls in four other schools, giving set F. ( The notation follows that used in other work with the same data.)

The data of set A are taken as the standard. To match the differing numbers of candidates in the two groups, the frequencies obtained from set A must first be multiplied by 97/105 ; the rest of the calculation proceeds as before. The items have been placed in order of difficulty, i.e., we are comparing answer-patterns. These patterns are graphed on the next page, and the calculation follows on the subsequent page.

In the table the frequencies of correct answers by the candidates of Set A are denoted $n_A$ ; that frequency multiplied by 97/105 is denoted by $n'_A$ ; and the frequencies from Set F are denoted by $n_F$ .

Fig. 8. Answer-patterns of M.H.T. 9.

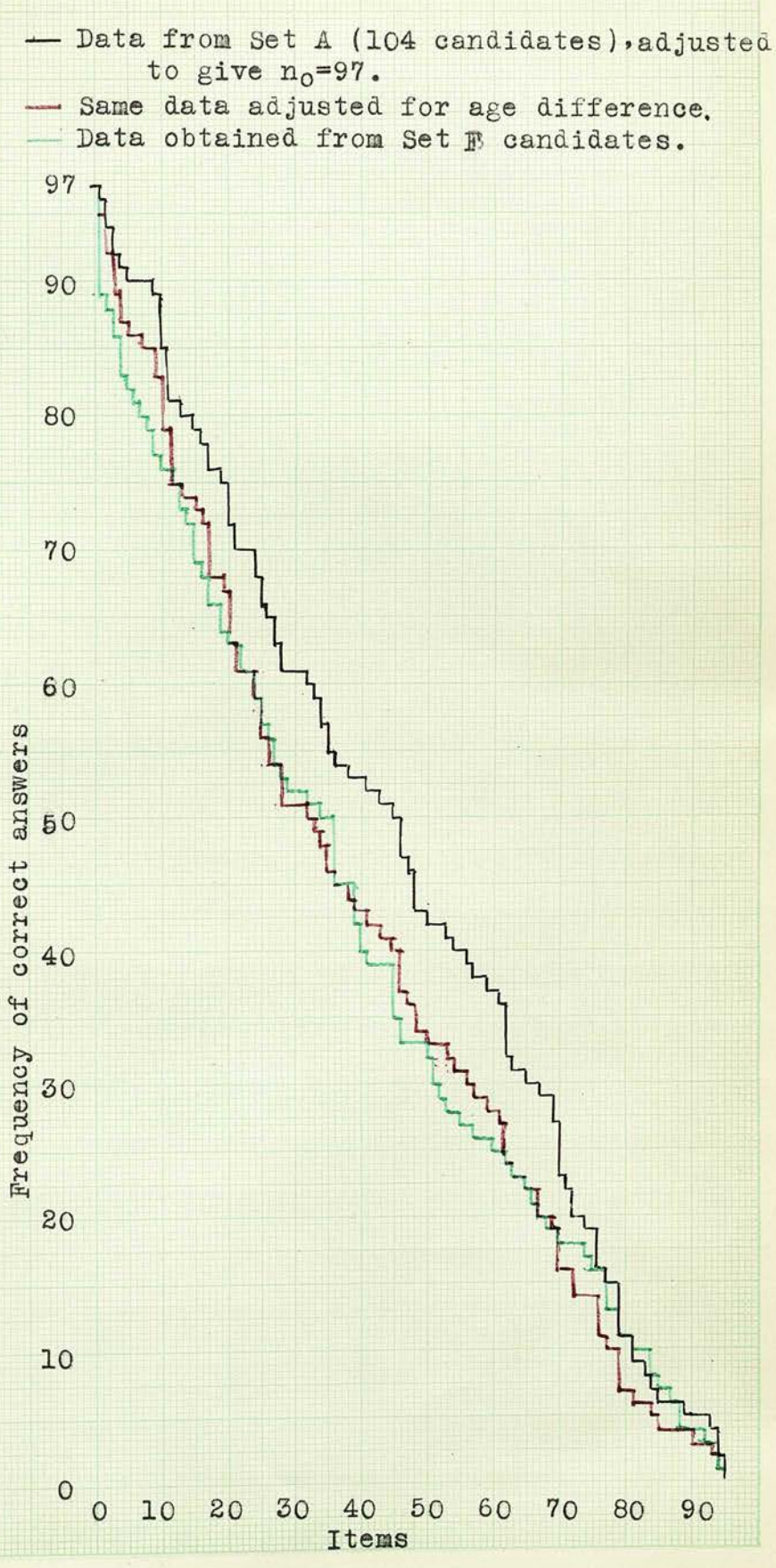


Table 7. Goodness of fit of M.H.T. 9 answer-patterns.

| $n_A$ | $n_A'$ | $n_F$ | $\frac{(n_A'-n_F)^2}{n_A'}$ | $n_A$ | $n_A'$ | $n_F$ | $\frac{(n_A'-n_F)^2}{n_A'}$ |
|---|---|---|---|---|---|---|---|
| 104 | 96 | 89 | .5 | 47 | 43 | 33 | 2.3 |
| 102 | 94 | 88 | .4 | 46 | 42 | 32 | 2.3 |
| 100 | 92 | 86 | .4 | 46 | 42 | 30 | 3.4 |
| 99 | 91 | 83 | .7 | 46 | 42 | 29 | 4.0 |
| 98 | 90 | 82 | .7 | 45 | 41 | 28 | 4.0 |
| 98 | 90 | 81 | .9 | 43 | 40 | 28 | 3.6 |
| 97 | 90 | 80 | 1.1 | 43 | 40 | 27 | 4.2 |
| 97 | 90 | 79 | 1.3 | 42 | 39 | 27 | 3.6 |
| 96 | 89 | 77 | 1.6 | 41 | 38 | 26 | 3.7 |
| 92 | 85 | 76 | 1.0 | 41 | 38 | 26 | 3.7 |
| 88 | 81 | 76 | .3 | 40 | 37 | 26 | 3.3 |
| 88 | 81 | 75 | .4 | 40 | 37 | 25 | 3.8 |
| 87 | 80 | 73 | .6 | 39 | 36 | 25 | 3.4 |
| 87 | 80 | 72 | .8 | 35 | 32 | 24 | 2.0 |
| 86 | 79 | 69 | 1.3 | 34 | 31 | 23 | 2.0 |
| 85 | 78 | 68 | 1.3 | 34 | 31 | 23 | 2.0 |
| 82 | 76 | 66 | 1.3 | 33 | 30 | 22 | 2.1 |
| 82 | 76 | 66 | 1.3 | 33 | 30 | 21 | 2.7 |
| 81 | 75 | 64 | 1.6 | 31 | 29 | 20 | 2.4 |
| 78 | 72 | 63 | 1.1 | 31 | 29 | 19 | 3.4 |
| 76 | 70 | 63 | .7 | 29 | 27 | 19 | 2.4 |
| 76 | 70 | 61 | 1.2 | 25 | 23 | 18 | 1.1 |
| 76 | 70 | 61 | 1.2 | 24 | 22 | 18 | .7 |
| 74 | 68 | 59 | 1.2 | 22 | 20 | 18 | .2 |
| 71 | 66 | 57 | 1.2 | 22 | 20 | 18 | .2 |
| 70 | 65 | 56 | 1.3 | 21 | 19 | 17 | .2 |
| 68 | 63 | 54 | 1.3 | 21 | 19 | 16 | .5 |
| 66 | 61 | 53 | 1.0 | 17 | 16 | 16 | 0 |
| 66 | 61 | 52 | 1.3 | 16 | 15 | 13 | .3 |
| 66 | 61 | 52 | 1.3 | 16 | 15 | 13 | .3 |
| 66 | 61 | 52 | 1.3 | 12 | 11 | 11 | 0 |
| 65 | 60 | 51 | 1.3 | 12 | 11 | 11 | 0 |
| 64 | 59 | 51 | 1.1 | 10 | 9 | 10 | .1 |
| 62 | 57 | 50 | .9 | 10 | 9 | 10 | .1 |
| 60 | 55 | 50 | .5 | 9 | 8 | 10 | .5 |
| 59 | 54 | 45 | 1.5 | 8 | 7 | 8 | .1 |
| 59 | 54 | 45 | 1.5 | 7 | 6 | 7 | .1 |
| 58 | 53 | 45 | 1.2 | 7 | 6 | 7 | .1 |
| 57 | 53 | 42 | 2.3 | 7 | 6 | 6 | 0 |
| 57 | 53 | 40 | 3.2 | 7 | 6 | 4 | .7 |
| 56 | 52 | 39 | 3.2 | 6 | 5 | 4 | .2 |
| 56 | 52 | 39 | 3.2 | 5 | 9 | 4 | .4 |
| 55 | 51 | 39 | 2.8 | 5 | 10 | 3 | 7 |
| 55 | 51 | 39 | 2.8 | 5 | | 3 | |
| 54 | 50 | 35 | 4.5 | 4 | 11 | 3 | 7 .8 |
| 51 | 47 | 33 | 4.2 | 2 | 10 | 1 | |
| 50 | 46 | 33 | 3.7 | | | | |
| 47 | 43 | 33 | 2.3 | 4686 | 4317 | 3624 | 142.7 |

Thus $\chi^2 = 142.7$ and the number of degrees of freedom, n, is 91.

$\sqrt{2\chi^2} - \sqrt{2n - 1} = 16.9 - 13.5 = 3.4$ .

The two answer-patterns are therefore significantly different. One cause of the difference is obvious from an inspection of the figures of the answer-patterns or from the graphs on page 52. The two groups are evidently not of the same average ability; since this is an intelligence test, probably it is directly due to the groups not being of the same age.

Is it possible to adjust the answer-pattern of Set A for this difference in average ability ? It cannot be done merely by subtracting a constant figure $\frac{4317-3624}{91}$ from each of the frequencies so as to make the total number of correct answers the same in both tests. It seems that the problem is soluble in the present case on the following assumptions:

(1) that each group has its ability normally distributed about a mental age equal to the chronological age of the group ( who are all of nearly of the same chronological age ), and that the standard deviation of their Intelligence Quotients is 13 points:

(2) that the number of correct answers to any question equals the number of candidates above a certain mental age:

(3) that the chronological ages of the two groups are known.

Under these conditions the answer-pattern may be adjusted by the following process.

Let the mean mental age of the first group be $A_1$ years, and that of the second group be $A_2$ years. Since the standard

deviation of I.Q.'s is 13 points, the standard deviations of the mental ages of the groups are $0.13A_1$ years and $0.13A_2$ years. For the groups considered $A_1$ is not very different from $A_2$, so that we may approximate, and use the same standard deviation for both distributions, taking it as $\frac{0.13}{2}(A_1+A_2)$ years. The difference of the mean mental ages, measured in $\sigma$ units, is therefore $\dfrac{A_1 - A_2}{\frac{0.13}{2}(A_1 + A_2)} = \delta$ say.

The following diagram represents the position and makes more clear the method used. Curve (A) represents the distribution of mental ages in group A; Curve (F) represents the distribution in group F. The line XY represents an item of such difficulty that it is answered correctly by the candidates of group A represented by the area of curve (A) to the right of XY. The number of correct answers to this item by group F is represented in the same way by the area of curve (F) to the right of XY.

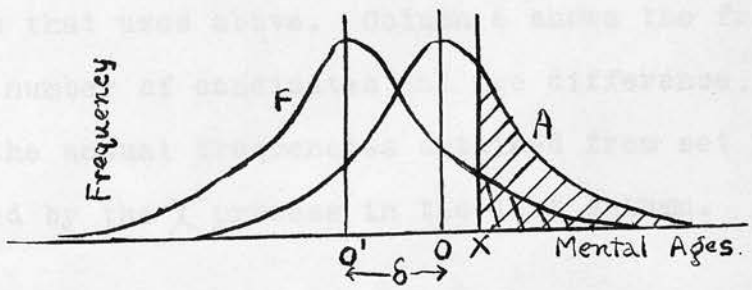



Fig. 9. Distribution of mental ages in groups A and F.

The steps to be taken with each item are then;

(1) from the number of correct answers by group A, find the area to the right of XY in curve A ; if the total area enclosed by

the curve and the x axis be taken as unity, this area equals $n_A/n_o$ :

(2) from tables of the probability integral find OX, the corresponding abscissa:

(3) by adding $\delta$, find O'X:

(4) from O'X, using the tables again, find the area to the right of XY in curve F.

(5) by multiplying by the number of candidates in group F, find the frequency of correct answers to the given item.($n'$)

Unfortunately, the average ages of the two groups tested were not known, so that $\delta$ could not be calculated directly. By an examination of the answer-pattern it is possible to estimate a value for $\delta$, which in the subsequent calculation is taken as 0.25 . This would be the case if the age of group A were 11 years 2 months and that of group F were 10 years 10 months.

Thus the following table was obtained. The notation used follows that used above. Column 6 shows the frequencies adjusted for number of candidates and age difference, and column 7 shows the actual frequencies obtained from set F. These are compared by the $\chi^2$ process in the last column.

Table 8. Adjustment of answer-pattern for age difference.

| $n_A$ | $n_A/n_o$ | OX | O'X | $n'/n_o$ | $n'$ | $n_E$ | $\frac{(n'-n_E)^2}{n'}$ |
|---|---|---|---|---|---|---|---|
| 104 | .99 | -2.33 | -2.08 | .98 | 95 | 89 | .4 |
| 102 | .97 | -1.88 | -1.63 | .95 | 92 | 88 | .2 |
| 100 | .95 | -1.64 | -1.39 | .92 | 89 | 86 | .1 |
| 99 | .94 | -1.55 | -1.30 | .90 | 87 | 83 | .2 |
| 98 | .93 | -1.48 | -1.23 | .89 | 86 | 82 | .2 |
| 98 | .93 | -1.48 | -1.23 | .89 | 86 | 81 | .3 |
| 97 | .92 | -1.41 | -1.16 | .88 | 85 | 80 | .3 |
| 97 | .92 | -1.41 | -1.16 | .88 | 85 | 79 | .4 |
| 96 | .91 | -1.34 | -1.09 | .86 | 83 | 77 | .4 |
| 92 | .87 | -1.13 | -0.88 | .81 | 79 | 76 | .1 |
| 88 | .84 | -0.99 | -0.74 | .77 | 75 | 76 | 0 |
| 88 | .84 | -0.99 | -0.74 | .77 | 75 | 75 | 0 |
| 87 | .83 | -0.95 | -0.70 | .76 | 74 | 73 | 0 |
| 87 | .83 | -0.95 | -0.70 | .76 | 74 | 72 | .1 |
| 86 | .82 | -0.92 | -0.68 | .75 | 73 | 69 | .2 |
| 85 | .81 | -0.88 | -0.63 | .74 | 72 | 68 | .2 |
| 82 | .78 | -0.77 | -0.52 | .70 | 68 | 66 | .1 |
| 82 | .78 | -0.77 | -0.52 | .70 | 68 | 66 | .1 |
| 81 | .77 | -0.74 | -0.49 | .69 | 67 | 64 | .1 |
| 78 | .74 | -0.64 | -0.39 | .65 | 63 | 63 | 0 |
| 76 | .72 | -0.58 | -0.33 | .63 | 61 | 63 | .1 |
| 76 | .72 | -0.58 | -0.33 | .63 | 61 | 61 | 0 |
| 76 | .72 | -0.58 | -0.33 | .63 | 61 | 61 | 0 |
| 74 | .70 | -0.52 | -0.27 | .61 | 59 | 59 | 0 |
| 71 | .67 | -0.44 | -0.19 | .58 | 56 | 57 | 0 |
| 70 | .66 | -0.41 | -0.16 | .56 | 54 | 56 | .1 |
| 68 | .65 | -0.39 | -0.14 | .56 | 54 | 54 | 0 |
| 66 | .63 | -0.33 | -0.08 | .53 | 51 | 53 | .1 |
| 66 | .63 | -0.33 | -0.08 | .53 | 51 | 52 | 0 |
| 66 | .63 | -0.33 | -0.08 | .53 | 51 | 52 | 0 |
| 66 | .63 | -0.33 | -0.08 | .53 | 51 | 52 | 0 |
| 65 | .62 | -0.31 | -0.06 | .52 | 50 | 51 | 0 |
| 64 | .61 | -0.28 | -0.03 | .51 | 49 | 51 | .1 |
| 62 | .59 | -0.23 | +0.02 | .49 | 48 | 50 | .1 |
| 60 | .57 | -0.18 | +0.07 | .47 | 46 | 50 | .3 |
| 59 | .56 | -0.15 | +0.10 | .46 | 45 | 45 | 0 |
| 59 | .56 | -0.15 | +0.10 | .46 | 45 | 45 | 0 |
| 58 | .55 | -0.13 | +0.12 | .45 | 44 | 45 | 0 |
| 57 | .54 | -0.10 | +0.15 | .44 | 43 | 42 | 0 |
| 57 | .54 | -0.10 | +0.15 | .44 | 43 | 40 | .2 |
| 56 | .53 | -0.08 | +0.17 | .43 | 42 | 39 | .2 |
| 56 | .53 | -0.08 | +0.17 | .43 | 42 | 39 | .2 |
| 55 | .52 | -0.05 | +0.20 | .42 | 41 | 39 | .1 |
| 55 | .52 | -0.05 | +0.20 | .42 | 41 | 39 | .1 |
| 54 | .51 | -0.03 | +0.22 | .41 | 40 | 35 | .6 |
| 51 | .48 | +0.05 | +0.30 | .38 | 37 | 33 | .4 |
| 50 | .47 | +0.08 | +0.33 | .37 | 36 | 33 | .3 |

| $n_A$ | $n_A/n_o$ | OX | O'X | $n'/n_o$ | n' | | $n_F$ | | $\frac{(n'-n_F)^2}{n'}$ |
|---|---|---|---|---|---|---|---|---|---|
| 47 | .45 | +0.13 | +0.38 | .35 | 34 | | 33 | | 0 |
| 47 | .45 | +0.13 | +0.38 | .35 | 34 | | 33 | | 0 |
| 46 | .44 | +0.15 | +0.40 | .34 | 33 | | 32 | | 0 |
| 46 | .44 | +0.15 | +0.40 | .34 | 33 | | 30 | | .3 |
| 46 | .44 | +0.15 | +0.40 | .34 | 33 | | 29 | | .5 |
| 45 | .43 | +0.18 | +0.43 | .33 | 32 | | 28 | | .5 |
| 43 | .41 | +0.23 | +0.48 | .32 | 31 | | 28 | | .3 |
| 43 | .41 | +0.23 | +0.48 | .32 | 31 | | 27 | | .5 |
| 42 | .40 | +0.25 | +0.50 | .31 | 30 | | 27 | | .3 |
| 41 | .39 | +0.28 | +0.53 | .30 | 29 | | 26 | | .3 |
| 41 | .39 | +0.28 | +0.53 | .30 | 29 | | 26 | | .3 |
| 40 | .38 | +0.31 | +0.56 | .29 | 28 | | 26 | | .2 |
| 40 | .38 | +0.31 | +0.56 | .29 | 28 | | 25 | | .3 |
| 39 | .37 | +0.33 | +0.58 | .28 | 27 | | 25 | | .2 |
| 35 | .33 | +0.44 | +0.69 | .25 | 24 | | 24 | | 0 |
| 34 | .32 | +0.47 | +0.72 | .24 | 23 | | 23 | | 0 |
| 34 | .32 | +0.47 | +0.72 | .24 | 23 | | 23 | | 0 |
| 33 | .31 | +0.50 | +0.75 | .23 | 22 | | 22 | | 0 |
| 33 | .31 | +0.50 | +0.75 | .23 | 22 | | 21 | | 0 |
| 31 | .29 | +0.55 | +0.80 | .21 | 20 | | 20 | | 0 |
| 31 | .29 | +0.55 | +0.80 | .21 | 20 | | 19 | | .1 |
| 29 | .28 | +0.58 | +0.83 | .20 | 19 | | 19 | | 0 |
| 25 | .24 | +0.71 | +0.96 | .17 | 16 | | 18 | | .3 |
| 24 | .23 | +0.74 | +0.99 | .16 | 16 | | 18 | | .3 |
| 22 | .21 | +0.81 | +1.06 | .14 | 14 | | 18 | | 1.1 |
| 22 | .21 | +0.81 | +1.06 | .14 | 14 | | 18 | | 1.1 |
| 21 | .20 | +0.84 | +1.09 | .14 | 14 | | 17 | | .6 |
| 21 | .20 | +0.84 | +1.09 | .14 | 14 | | 16 | | .3 |
| 17 | .16 | +0.99 | +1.24 | .11 | 11 | | 16 | | 2.3 |
| 16 | .15 | +1.04 | +1.29 | .10 | 10 | | 13 | | .9 |
| 16 | .15 | +1.04 | +1.29 | .10 | 10 | | 13 | | .9 |
| 12 | .11 | +1.23 | +1.48 | .07 | 7 | | 11 | | 2.3 |
| 12 | .11 | +1.23 | +1.48 | .07 | 7 | | 11 | | 2.3 |
| 10 | .09 | +1.34 | +1.59 | .06 | 6 | | 10 | | 2.7 |
| 10 | .09 | +1.34 | +1.59 | .06 | 6 | | 10 | | 2.7 |
| 9 | .09 | +1.34 | +1.59 | .06 | 6 | | 10 | | 2.7 |
| 8 | .08 | +1.41 | +1.66 | .05 | 5 | | 8 | | 1.8 |
| 7 | .07 | +1.48 | +1.73 | .04 | 4 | | 7 | | |
| 7 | .07 | +1.48 | +1.73 | .04 | 4 | 12 | 7 | 20 | 5.3 |
| 7 | .07 | +1.48 | +1.73 | .04 | 4 | | 6 | | |
| 7 | .07 | +1.48 | +1.73 | .04 | 4 | | 4 | | |
| | | | | | 4 | 8 | 4 | 8 | 0 |
| 6 | .06 | +1.55 | +1.80 | .04 | | | | | |
| 5 | .05 | +1.64 | +1.89 | .03 | 3 | 6 | 4 | 7 | |
| 5 | .05 | +1.64 | +1.89 | .03 | 3 | | 3 | | .2 |
| 5 | .05 | +1.64 | +1.89 | .03 | 3 | | 3 | | |
| 4 | .04 | +1.75 | +2.00 | .02 | 2 | 6 | 3 | 7 | .2 |
| 2 | .02 | +2.05 | +2.30 | .01 | 1 | | 1 | | |
| | | | | | 3661 | | 3624 | | 38.1 |

This has certainly achieved the equalisation of average ability between the hypothetical and observed answer-patterns, as shown by the values of $\Sigma n$, 3661 and 3624. The value of $\Sigma n$ for the original form of the Set A answer-pattern, adjusted only for the different number of candidates, was 4317. The fit of the two patterns is also very much better. The value of $\chi^2$ is now only 38.1 , so low a value as almost to be suspicious, since for 88 degrees of freedom a lower value of $\chi^2$ would occur in only one case in about 200,000 trials. ( P = .999995 )

We may therefore conclude that any difference between the answer-patterns of Sets A and F is due to the age difference, and not to any lack of permanence in the pattern itself.

Example 5. The preceding examples have established the permanence of an answer-pattern when the same test is given to a similar population. The following example deals with a more complicated case. If from any test a number of items are selected and put together to form a new test, does the answer-pattern obtained from the old test still hold under the new conditions, given a similar population ? This question can be answered only by experimenting; the results of an experiment of this type will now be examined.

The tests A, B, and C employed in the author's B. Ed. thesis were composed of items chosen from M.H.T. 9. The tests were constructed to provide certain types of answer-pattern, the basis of selection being the answer-patterns provided by the

202 candidates of groups A and F already mentioned.

Unfortunately when the tests were applied, the candidates now attempting them were younger, and in some cases much younger than the M.H.T. groups. Part of the data had to be rejected for this reason, the age disparity being too great, but use was made of the remaining papers, the work of 166 candidates of an average age 10 years 6 months, which was 6 months less than the average age of the M.H.T. groups.

In the comparison of the answer-patterns, the first step must then be the adjusting of the M.H.T. pattern for age difference. Proceeding as before,

$$\delta = \frac{A_1 - A_2}{\frac{.13}{2}(A_1 + A_2)} = .36$$

The adjustment of the answer-pattern for number of candidates and age difference then proceeds as before. In the tables shown n denotes frequency of correct answers by 202 candidates in M.H.T. 9 ; n' denotes this frequency adjusted for number of candidates and age difference ; and n'' denotes the frequency of correct answers actually obtained from the 166 candidates. All three tests were 15 item tests, and the results are shown without any rearrangement of the order of items. The first test to be considered is test B.

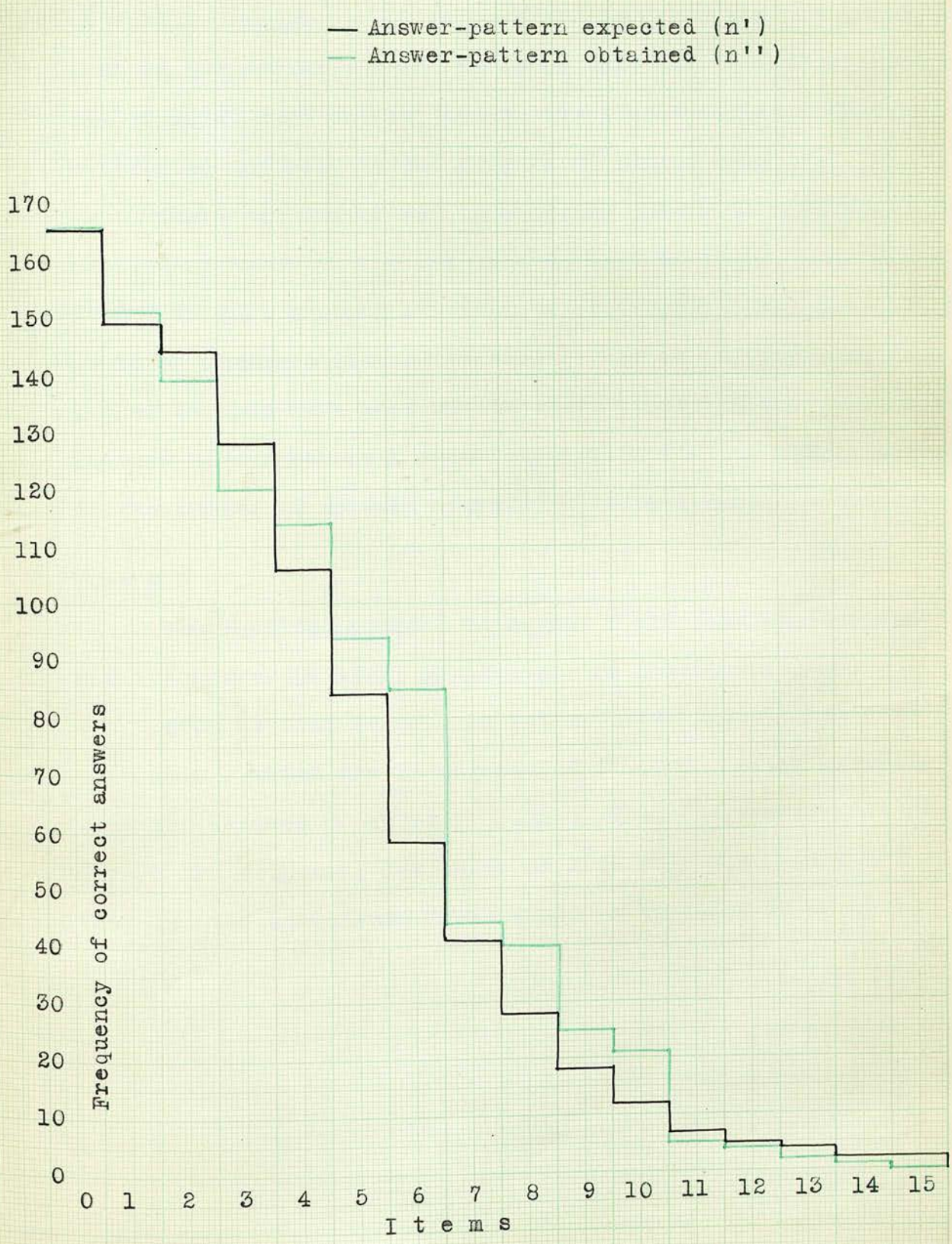
Table 9. Comparison of answer-patterns intended (n') and obtained (n'') for test B.

| Item | n (M.H.T.9) | n'(adjusted) | n'' |
|---|---|---|---|
| 0 | 202 | 166 | 166 |
| 1 | 192 | 149 | 120 |
| 2 | 188 | 144 | 151 |
| 3 | 175 | 128 | 139 |
| 4 | 155 | 106 | 114 |
| 5 | 130 | 84 | 85 |
| 6 | 99 | 58 | 40 |
| 7 | 75 | 41 | 94 |
| 8 | 56 | 28 | 25 |
| 9 | 39 | 18 | 21 |
| 10 | 27 | 12 | 44 |
| 11 | 19 | 7 | 5 |
| 12 | 14 | 5 | 4 |
| 13 | 11 | 4 | 1 |
| 14 | 7 | 2 | 2 |
| 15 | 5 | 2 | 0 |
| | | 788 | 845 |

The two answer-patterns are graphed overleaf.

When the items have been rearranged to form the answer-pattern, and the $\chi^2$ method has been applied, it is found that

Fig. 10. Answer-patterns of Thesis Test B.



$\chi^2 = 33.1$ , which for 12 degrees of freedom gives P = .0003, that is the deviations are so large that they would occur in only one of about 3000 trials, and so the discrepancies cannot be attributed entirely to the effects of random sampling.

There are one or two interesting points to note about the changes apparent in the response to certain items. For example, item 7 in test B stood at the head of a page. Like item 10, it could be answered without referring back to the previous page, as had to be done with most of the other items. These favourable factors have apparently been responsible for the marked increase in the number of correct answers to these items.

Test C

The results obtained with test C are tabulated overleaf, and the answer-patterns graphed on the subsequent page.

Even as they stand, the fit of these distributions is good. $\chi^2 = 17.0$ , which for 14 degrees of freedom ( after grouping the last two classes ) gives P = .25 . When the answer-patterns are formed and compared, the value of $\chi^2$ is only 8.4 , giving P = .86 ; that is, the deviations found would be exceeded in 86 of 100 trials.

Table 10. Comparison of answer-patterns intended (n') and obtained (n'') for test C.

| Item | n (M.H.T.9) | n'(adjusted) | n'' |
|---|---|---|---|
| 0 | 202 | 166 | 166 |
| 1 | 180 | 134 | 142 |
| 2 | 179 | 133 | 127 |
| 3 | 177 | 130 | 121 |
| 4 | 176 | 129 | 117 |
| 5 | 174 | 127 | 132 |
| 6 | 167 | 120 | 115 |
| 7 | 160 | 112 | 105 |
| 8 | 146 | 98 | 80 |
| 9 | 129 | 83 | 72 |
| 10 | 111 | 68 | 55 |
| 11 | 89 | 51 | 68 |
| 12 | 62 | 32 | 37 |
| 13 | 36 | 17 | 18 |
| 14 | 16 | 7 | 5 |
| 15 | 6 | 2 | 4 |
| | | 1243 | 1198 |

Fig. 11. Answer-patterns of Thesis Test C.

—Answer-pattern expected (n')
—Answer-pattern obtained (n'')

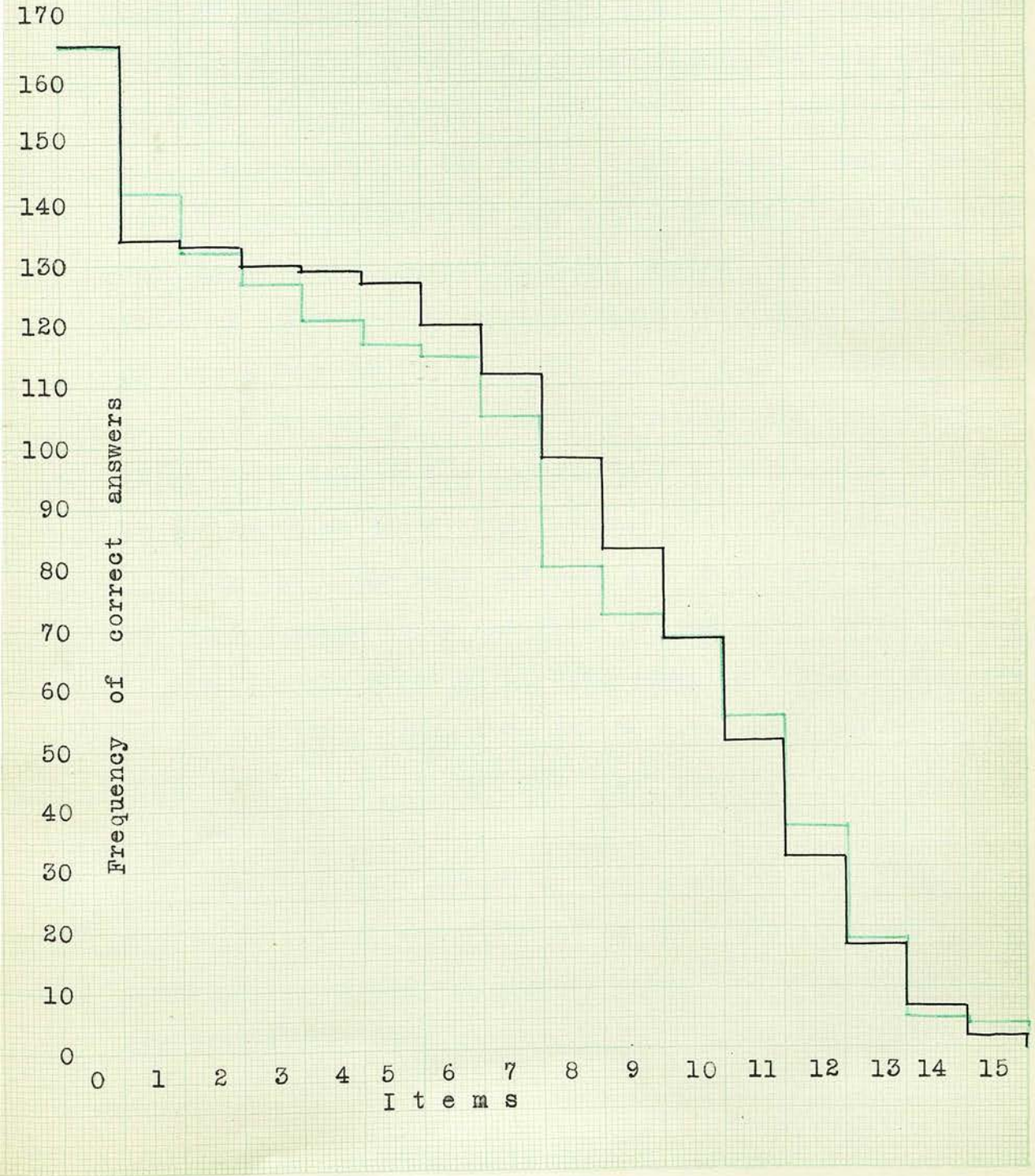


Table 11. Comparison of frequencies of correct answers expected (n') and obtained (n'') for test A.

| Item | n (M.H.T.9) | n'(adjusted) | n'' |
|---|---|---|---|
| 0 | 202 | 166 | 166 |
| 1 | 81 | 45 | 32 |
| 2 | 82 | 46 | 30 |
| 3 | 90 | 51 | 35 |
| 4 | 92 | 53 | 24 |
| 5 | 93 | 53 | 21 |
| 6 | 96 | 56 | 20 |
| 7 | 97 | 56 | 18 |
| 8 | 98 | 57 | 54 |
| 9 | 102 | 61 | 63 |
| 10 | 103 | 61 | 43 |
| 11 | 106 | 63 | 29 |
| 12 | 112 | 69 | 51 |
| 13 | 113 | 69 | 61 |
| 14 | 117 | 72 | 57 |
| 15 | 118 | 74 | 17 |
| | | 886 | 535 |

Even when the items are arranged in order of difficulty, $\chi^2 = 176$ which for 15 degrees of freedom gives a probability of much less than one in a million that the deviations are due to random sampling.

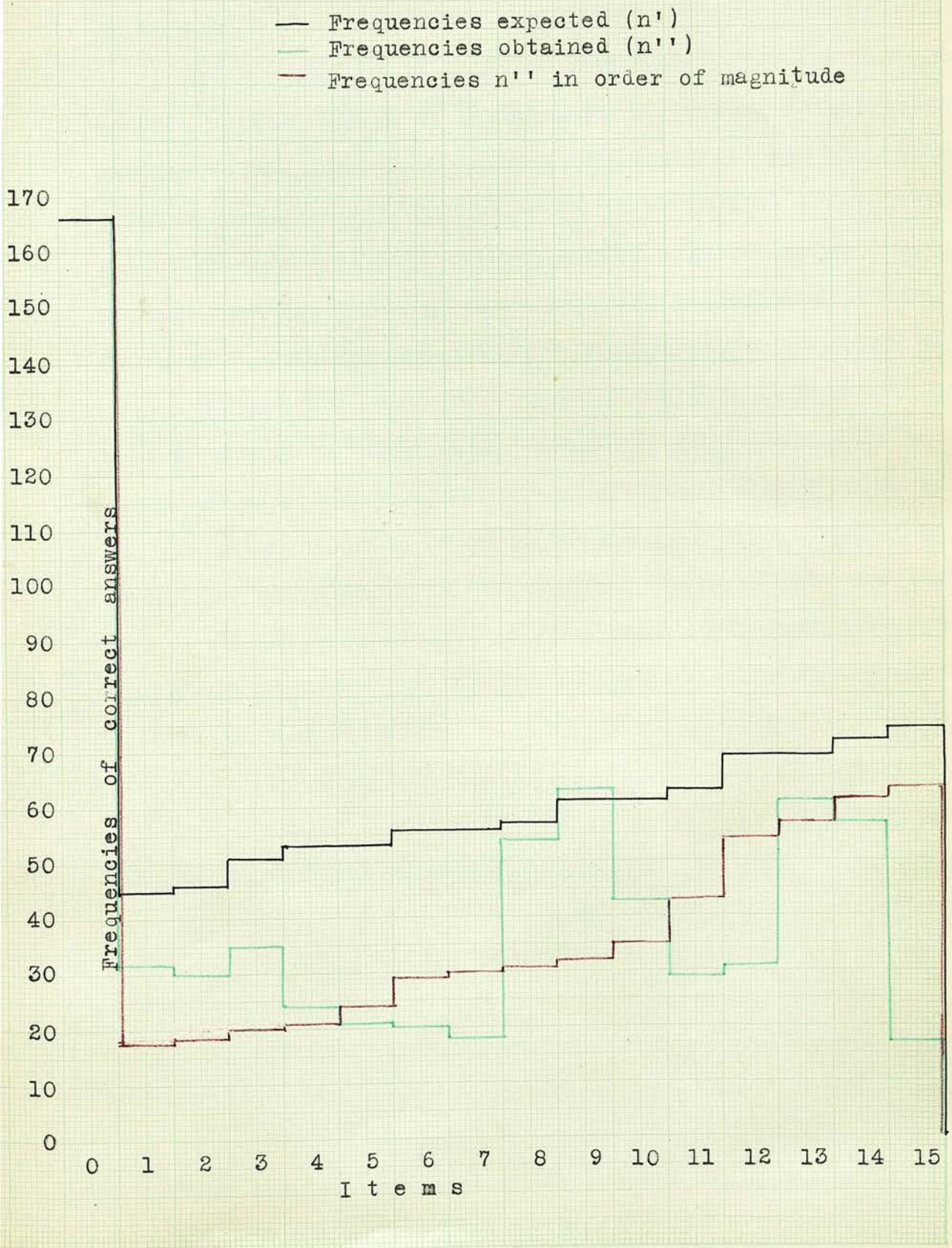
Test A.

Test A was in several ways a peculiar test. Its items were arranged in decreasing order of difficulty, the hardest being first, and there was more reading than usual preceding each item. None of the items was very easy, even for the original group of testees, and the effect of the age difference made them more difficult still. Although at the time the fact was not recognised, the reversal of the order of the items probably added yet more to the difficulty of the test. It has since been pointed out, in an investigation by E. S. Souter, that reversing the order of difficulty in similar tests increased the percentage of errors from 33.35% to 41.75% in the case of spelling, from 33.07% to 38.81% in the case of problems, and from 9.71% to 13.82% in a test of the fundamentals of arithmetic. ( Published in the Scottish Educational Journal, 24th May 1935.).

On the other hand, it must not be thought that test A suffered through being the first of the three tests to be given. To equalise the practice effects among the three tests, one sixth of the candidates took them in the order ABC, one sixth in the order ACB, and so through the other possible orders BAC, BCA, CAB, CBA.

The resultant effect on test A was to make it much more difficult than it was intended to be. This is shown by the total number of correct answers being 535, as against the expected number 886.

Fig. 12. Frequencies of correct answers in Thesis Test A



Conclusion.

These results show that the answer-pattern of a test does possess a large degree of permanence, if certain rather obvious conditions are observed. If a pattern derived from the answers given by one group of candidates is to be used to predict the answer-pattern that will be produced by a second group, then the age (or ability) difference between the groups must be small. For a small difference the answer-pattern may be adjusted by the method illustrated, but for a large difference this method has been proved to be inadequate. Secondly, if a test is to be constructed from some of the items of a test with a known answer-pattern, then the test-paper so constructed should have a format as similar as possible to the original test, so that the items have a similar environment. Under these conditions, it seems highly probable that the answer-pattern obtained will correspond closely to that intended.

## Chapter 5. The Relation of Answer-pattern and Score-scatter in the General Case.

In the special case considered in chapter 2, there is an exact relationship between the answer-pattern and the score-scatter, expressed by the equations A and B of that chapter. The necessary condition was that each candidate's score be made up of answers to the easiest questions, and this state of affairs was described as unig. Tests of this type are very rare; a certain amount of hig enters into the composition of almost every test, and the exactness of the relationship between answer-pattern and score-scatter is destroyed. Although the exact relationship has gone, it may be that there persists some measure of correlation, or, to express the idea in a different way, there may stll be some degree of control of the score-scatter by the answer-pattern.

As examples, consider the following answer-patterns with their answer-pattern-differentials and score-scatters, as obtained in actual examinations. The first two are selected more or less at random from the 41 tests, of which they are tests 32 and 7, and the third from data of M.H.T. 12p already used.

Table 12. Data of test 32, test 7, and M.H.T. 12p .

Test 32

| Item or score. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Answer-pattern. | 32 | 31 | 30 | 30 | 27 | 27 | 21 | 19 | 19 | 4 | 2 |
| Answer-pattern-differential. | 1 | 1 | 0 | 3 | 0 | 6 | 2 | 0 | 15 | 2 | 2 |
| Score-scatter. | 0 | 0 | 0 | 0 | 3 | 7 | 3 | 11 | 5 | 2 | 1 |

Test 7.

| Item or score. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Answer-pattern. | 32 | 32 | 27 | 20 | 18 | 17 | 17 | 17 | 14 | 13 | 11 |
| Answer-pattern-differential. | 0 | 5 | 7 | 2 | 1 | 0 | 0 | 3 | 1 | 2 | 11 |
| Score-scatter. | 0 | 1 | 3 | 4 | 5 | 3 | 2 | 1 | 7 | 3 | 3 |

M.H.T. 12p.

| Item or score. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Answer-pattern. | 450 | 431 | 416 | 402 | 399 | 389 | 386 | 268 | 205 | 179 |
| Answer-pattern-differential. | 19 | 15 | 14 | 3 | 10 | 3 | 118 | 63 | 26 | 179 |
| Score-scatter. | 10 | 9 | 11 | 13 | 14 | 25 | 69 | 97 | 108 | 94 |

It will be readily seen that corresponding frequencies in these tables are far from being equal. Pearson's test of goodness of fit would at once declare the answer-pattern-differential and the score-scatter to be different distributions; there is not the faintest possibility that the deviations are due to errors of sampling. Yet the data seem to indicate that there is still some control of the score-scatter by the answer-pattern-differential.

For example, the two-humped answer-pattern-differential of test 7 is accompanied by a distinctly double-humped score-scatter. The positively skewed answer-pattern-differential of M.H.T. 12p is accompanied by a positively skewed score-scatter.

There are two parameters of the two distributions that are necessarily identical. The total frequency in both answer-pattern -differential and score-scatter is the same; it is $n_o$, the number of candidates. Also the first moment of both distributions is the same. In the case of the answer-pattern-differential it is $\sum_{o}^{m} x(n_x - n_{x+1})$ which equals $n_1 + n_2 + n_3 + \ldots + n_m$; that is, the number of points scored. For the score-scatter the first moment is $\sum_{o}^{m} xN_x$, which again equals the number of points scored. Expressed in another way, these latter equations mean that both distributions have the same mean. This is quite independent of the presence or absence of hig.

This line of thought suggests that with a sufficient number of tests given to the same candidates we might correlate corresponding moments. Alternatively, instead of correlating the second moments we might correlate the standard deviations; and for the third moments calculation we might substitute correlation of skewness. It is hardly necessary to proceed further than the third moment for two reasons. First, the third is the highest moment with which an examiner is ordinarily concerned, the mean score being calculable from the first moment, the standard deviation from the second and first, and the skewness from the

first three moments. Second, it is a fact well known to statisticians that spurious deviations sometimes occur in an extensive set of observations; though their effect may be small on the first, second, or even third moments, it increases rapidly with higher moments and may outweigh the sum of all the other terms, so that the calculated moment becomes quite unreliable.

The correlation of standard deviations and of coefficients of skewness was one of the main lines of attack in this investigation. The correlation of the standard deviations of answer-pattern-differential and score-scatter is reported in chapter 6, and the correlation of the measures of skewness in chapter 7.

In the case of a single test, the estimation of the degree of relationship between the answer-pattern-differential and the score-scatter is not so easy. Several methods have been tried by the author, but none has proved entirely satisfactory. The subject will be more fully discussed and the coefficients devised will be described and criticised in chapter 8.

## Chapter 6. The Correlation of the Standard Deviations of Answer-pattern-differential and Score-scatter.

### 1. Theory of the method used.

The term standard deviation was first used in the study of normal frequency distributions. If a variate x is normally distributed then two statistics, the mean and the standard deviation, summarise our knowledge of the distribution. The standard deviation, $\sigma$, measures the extent to which the variate x is scattered about the mean a.

This theory may be applied readily enough to the score-scatters found in the tests, for they approximate to normal frequency distributions, but such is not the case with the answer-pattern-differentials. These latter belong to none of Pearson's types of frequency curves. However, it seems justifiable to apply the same method of calculation to find for the answer-pattern-differential a pseudo- standard deviation, on the ground that the statistic so calculated is to be used solely for measuring the spread or scatter of the observations from the mean; no property of standard deviations which applies only to their use in normal distributions is employed.

In partial support of this argument, there may be quoted the opinion of Bowley: "Even when a group has no very close connection with the first or second approximations to the curve of error, it seems probable that c ( $=\sqrt{2}\sigma$) and j ( $=\frac{1}{2\sqrt{2}}S$ ), calculated by the methods which for the particular group seem the

least liable to chance disturbance, are the best single measures of the groupings about the average and the skewness that we can devise." ( Elements of Statistics, page 331.)

2. Experimental Results.

Very suitable data for measuring the correlation of the two standard deviations are furnished by the 41 tests. They were attempted by the same candidates throughout, and had each 10 items, avoiding any heterogeneity due to differing numbers of items.

The method of calculation is shown below; the test chosen as an example is number 4 of the 41 tests. ( See table 13 overleaf)

A provisional mean is chosen at x = 4. The first and second moments about that mean are calculated; if these are den denoted by $m_1$ and $m_2$ it is easy to show that $\sigma^2 = m_2 - m_1^2$. A check is provided by the fact already mentioned that the two distributions have the same first moment. In addition all the results were checked by repeating the calculations with a fresh provisional mean.

Table 13. Standard deviations of answer-pattern-differential and score-scatter.

1. Answer-pattern-differential.

| Item | A.P. | A.P.D. | x | fx | $fx^2$ |
|---|---|---|---|---|---|
| 0 | 32 | 3 | -4 | -12 | 48 |
| 1 | 29 | 2 | -3 | - 6 | 18 |
| 2 | 27 | 4 | -2 | - 8 | 16 |
| 3 | 23 | 5 | -1 | - 5 | 5 |
| 4 | 18 | 6 | 0 | - 0 | 0 |
| 5 | 12 | 3 | 1 | 3 | 3 |
| 6 | 9 | 1 | 2 | 2 | 4 |
| 7 | 8 | 1 | 3 | 3 | 9 |
| 8 | 7 | 1 | 4 | 4 | 16 |
| 9 | 6 | 0 | 5 | 0 | 0 |
| 10 | 6 | 6 | 6 | 36 | 216 |
| Total marks.145 | | $n_o$=32 | | +17 | 335 |

$m_1 = 17/32 = +0.53$ $\quad m_2 = 335/32 = 10.47$

$$\sigma_{A.P.D.} = \sqrt{m_2 - m_1^2} = \underline{3.19}$$

2. Score-scatter.

| Item | N | x | Nx | $Nx^2$ |
|---|---|---|---|---|
| 0 | 1 | -4 | -4 | 16 |
| 1 | 1 | -3 | -3 | 9 |
| 2 | 1 | -2 | -2 | 4 |
| 3 | 10 | -1 | -10 | 10 |
| 4 | 5 | 0 | 0 | 0 |
| 5 | 4 | 1 | 4 | 4 |
| 6 | 4 | 2 | 8 | 16 |
| 7 | 2 | 3 | 6 | 18 |
| 8 | 2 | 4 | 8 | 32 |
| 9 | 2 | 5 | 10 | 50 |
| 10 | 0 | 6 | 0 | 0 |
| | $n_o$..32 | | +17 | 159 |

$m_1 = 17/32 = +0.53$ $\quad m_2 = 159/32 = 4.98$

$$\sigma_N = \sqrt{m_2 - m_1^2} = \underline{2.17}$$

Check: 32×4 + 17 = 145 = total marks scored, as in answer-pattern.

Table 14. Standard deviations of answer-pattern-differential and of score-scatter of the 41 tests.

| Test | $\sigma_{A.P.D.}$ | $\sigma_{N.}$ |
|---|---|---|
| 1 | 3.11 | 2.26 |
| 2 | 3.38 | 2.26 |
| 3 | 3.64 | 2.35 |
| 4 | 3.19 | 2.17 |
| 5 | 3.37 | 2.18 |
| 6 | 3.01 | 1.96 |
| 7 | 3.79 | 2.71 |
| 8 | 2.74 | 2.08 |
| 9 | 3.08 | 1.93 |
| 10 | 3.24 | 2.22 |
| 11 | 3.84 | 2.46 |
| 12 | 3.61 | 2.11 |
| 13 | 2.89 | 2.13 |
| 14 | 3.42 | 2.29 |
| 15 | 3.89 | 2.17 |
| 16 | 2.90 | 1.90 |
| 17 | 2.54 | 1.73 |
| 18 | 3.67 | 2.32 |
| 19 | 3.61 | 2.31 |
| 20 | 3.21 | 2.11 |
| 21 | 3.20 | 1.92 |
| 22 | 3.34 | 1.94 |
| 23 | 2.86 | 1.77 |
| 24 | 2.99 | 1.75 |
| 25 | 3.33 | 2.05 |
| 26 | 3.16 | 1.97 |
| 27 | 2.92 | 2.18 |
| 28 | 3.93 | 2.37 |
| 29 | 3.35 | 2.40 |
| 30 | 3.08 | 2.02 |
| 31 | 2.70 | 1.99 |
| 32 | 2.47 | 1.52 |
| 33 | 2.70 | 1.49 |
| 34 | 2.88 | 2.09 |
| 35 | 3.11 | 2.03 |
| 36 | 3.42 | 2.10 |
| 37 | 3.51 | 2.38 |
| 38 | 3.79 | 2.51 |
| 39 | 3.56 | 2.48 |
| 40 | 3.09 | 2.09 |
| 41 | 3.43 | 2.28 |
| Means | 3.25 | 2.12 |

Table 15. Correlation of standard deviations of answer-pattern-differential and score-scatter. $r = .789$.

$\sigma_{A.P.D.}$

| $\sigma_N$ | 2.4 to 2.5 | 2.5 to 2.6 | 2.6 to 2.7 | 2.7 to 2.8 | 2.8 to 2.9 | 2.9 to 3.0 | 3.0 to 3.1 | 3.1 to 3.2 | 3.2 to 3.3 | 3.3 to 3.4 | 3.4 to 3.5 | 3.5 to 3.6 | 3.6 to 3.7 | 3.7 to 3.8 | 3.8 to 3.9 | 3.9 to 4.0 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2.7 to 2.8 | | | | | | | | | | | | | | 1 | | | 1 |
| 2.6 to 2.7 | | | | | | | | | | | | | | | | | 0 |
| 2.5 to 2.6 | | | | | | | | | | | | | | 1 | | | 1 |
| 2.4 to 2.5 | | | | | | | | | | 1 | | 1 | | | 1 | | 3 |
| 2.3 to 2.4 | | | | | | | | | | | | 1 | 3 | | | 1 | 5 |
| 2.2 to 2.3 | | | | | | | | 1 | 1 | 1 | 2 | | | | | | 5 |
| 2.1 to 2.2 | | | | | 1 | 1 | | 1 | 1 | 1 | 1 | | 1 | | 1 | | 8 |
| 2.0 to 2.1 | | | | 1 | 1 | | 2 | 1 | | 1 | | | | | | | 6 |
| 1.9 to 2.0 | | | | 1 | | 1 | 2 | 1 | 1 | 1 | | | | | | | 7 |
| 1.8 to 1.9 | | | | | | | | | | | | | | | | | 0 |
| 1.7 to 1.8 | | 1 | | | 1 | 1 | | | | | | | | | | | 3 |
| 1.6 to 1.7 | | | | | | | | | | | | | | | | | 0 |
| 1.5 to 1.6 | 1 | | | | | | | | | | | | | | | | 1 |
| 1.4 to 1.5 | | | | 1 | | | | | | | | | | | | | 1 |
| Totals | 1 | 1 | 0 | 3 | 3 | 3 | 4 | 4 | 3 | 5 | 3 | 2 | 4 | 2 | 2 | 1 | 41 |

The correlation found from the preceding table is $r = .789$ , a fairly high positive value. The usual formula for the probable error of a correlation coefficient derived from n pairs of observations, i.e., $P.E. = \frac{.6745(1-r^2)}{\sqrt{n}}$ would give for the above value of r a probable error of $\pm .041$ . It has, however, been pointed out by Fisher that the use of this formula is often misleading, especially in a case like this where the sample is small, and the correlation coefficient high. The distribution of correlation coefficients is often far from normal with small samples, and even with large samples if the correlation is high.

The method proposed as an alternative is to transform the coefficient by the substitution

$$z = \tfrac{1}{2}\log_e \frac{1+r}{1-r} \quad .$$

The distribution of z tends to normality rapidly as the sample is increased, whether r is large or not. The standard deviation of z is $\frac{1}{\sqrt{n-3}}$ , depending only on the number of observations. In Fisher's "Statistical Methods" there is provided a table (Table VB) giving the value of z for any r.

For $r = .789$ we find $z = 1.069$

$$\sigma_z = \frac{1}{\sqrt{38}} = .162$$

The value of z found therefore differs significantly (i.e. by $2\sigma$) from $1.069 \pm .324$ , i.e. from $z = 1.393$ and $z = .745$ . The corresponding values of r are .884 and .632 . There is thus no

doubt

dubiety about the significance of the correlation found; it is almost certainly greater than .632 , and less than .884 .

The regression equation showing the regression of $\sigma_N$ on $\sigma_{A.P.D.}$ is

$$\sigma_N = 0.55\,\sigma_{A.P.D.} + 0.34 ,$$

where $\sigma_N$ and $\sigma_{A.P.D.}$ are measured in the same units as used in their calculation originally. This estimate has a fairly high reliability by reason of the high value of r.

For a given value of $\sigma_{A.P.D.}$ the array of $\sigma_N$ has a standard deviation $s' = s\sqrt{1 - r^2}$ , where s is the standard deviation of the total array of all the $\sigma_N$ 's . In the present case, s was found to equal $\sqrt{2.68/41}$ so that

$$s' = \sqrt{\frac{2.68}{41}(1 - .789^2)} = .16$$

For example, if $\sigma_{A.P.D.} = 3$, the most probable value of $\sigma_N$ would be 1.99 , and about two thirds of the values of $\sigma_N$ expected would fall in the range 1.83 to 2.15 .

## 3. Significance of the Results.

These results are based on assumptions of normality in the arrays, assumptions whose truth can hardly be tested with so sm small an amount of data. It must also be remembered that they have been derived from the 41 tests and are strictly applicable only to those 41 tests. At the same time their general trend is significant.

Consider for example an answer-pattern of the "flat" type, as illustrated in the figure below.

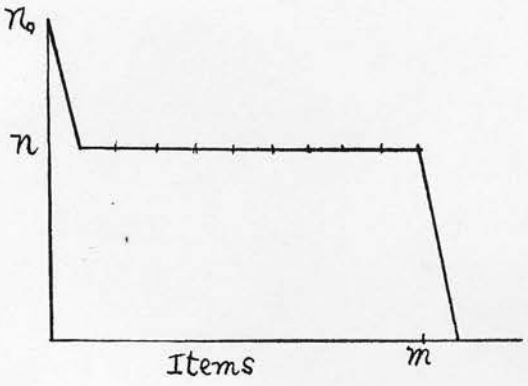



Figure 13. A flat answer-pattern.

The corresponding answer-pattern-differential is $n_o$-n, 0, 0, .....0, n. This has a large $\sigma_{A.P.D.}$ and therefore will probably produce a score-scatter with a large value of $\sigma_N$, that is the scores will be widely distributed about the mean.

Consider secondly a test of the "steep" type.

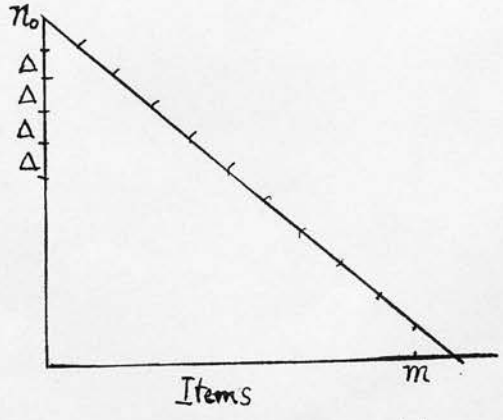



Figure 14. A steep answer-pattern.

This test has an answer-pattern-differential $\Delta, \Delta, \Delta, .., \Delta$, which has an intermediate value of $\sigma_{A.P.D.}$. The standard

deviation of the scaores in such a test would therefore be expected to be of average size for the number of items in the test.

From a test with a very small $\sigma_{A.P.D.}$, such as would be produced by the answer-pattern below, a very small $\sigma_{\pi}$ would be expected.

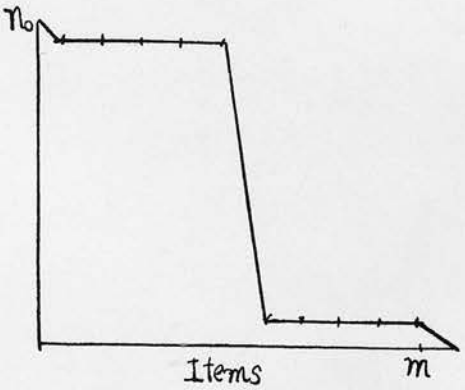



Figure 15. Answer-pattern producing small $\sigma_{A.P.D.}$.

Here the answer-pattern-differential is $n_0-n_1$, 0, 0,.....0, $n_1-n_2$, 0, 0,....,$n_2$. If $n_0-n_1$ and $n_2$ are small, the value of $\sigma_{A.P.D.}$ is small.

An attempt was made to apply the above methods of investigation to the complete tests, but difficulties arose through the varying numbers of items in each test, and through the effects of grouping in these tests which had many items. The number of these tests which are usable is also small.

4. Additional confirmatory data.

As a practical test of these principles, three tests were designed to follow roughly the lines shown in figures 13-15. They were 8-item physics tests and were given to 34 pupils in the

first year of a secondary school.

Test D was designed as a fairly steep test of the type shown in figure 14 ; test E was a two level answer-pattern of the type of figure 15 ;,and test F was a rather flat test somewhat similar to that in figure 13. The usual precautions were taken to neutralise practice and fatigue effects, the various cyclic orders of the tests being as far as possible equally used. The tests were typed on separate sheets. The results are given below in tabular form, and the answer-patterns and score-scatters are aet out in histogram form on pages 85-87.

Table 16. Answer-patterns and score-scatters of tests D,E,F.

Test D

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Answer-pattern | 34, | 34, | 27, | 22, | 21, | 18, | 12, | 8, | 5. |
| Score-scatter | 0, | 2, | 1, | 6, | 9, | 10, | 4, | 1, | 1. |

Standard deviation of A.P.D. 2.48

Standard deviation of score-scatter 1.49

Mean score 4.3

Test E

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Answer-pattern | 34, | 32, | 30, | 30, | 29, | 9, | 7, | 6, | 2. |
| Score-scatter | 0, | 0, | 4, | 2, | 17, | 6, | 3, | 1, | 1. |

Standard deviation of A.P.D. 1.90

Standard deviation of score-scatter 1.31

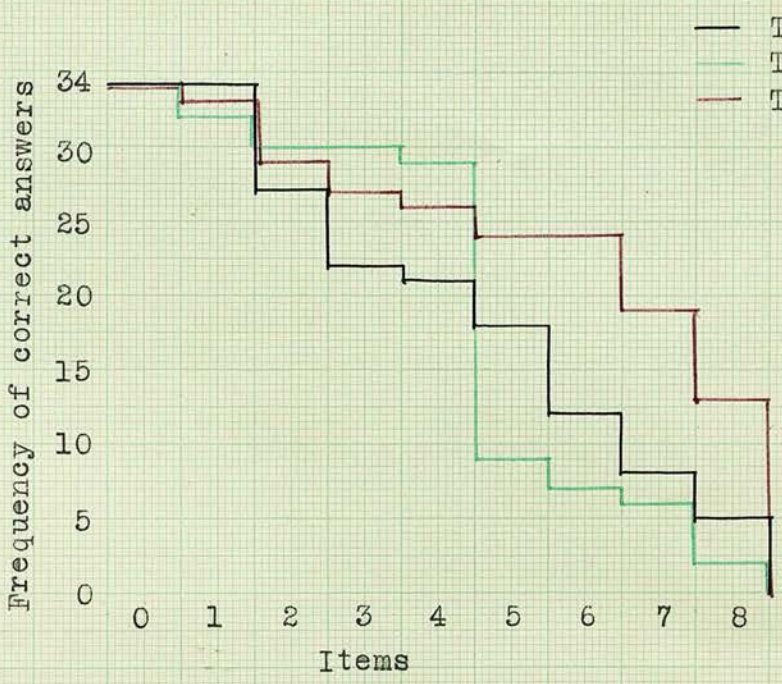
Mean score 4.3

Test F

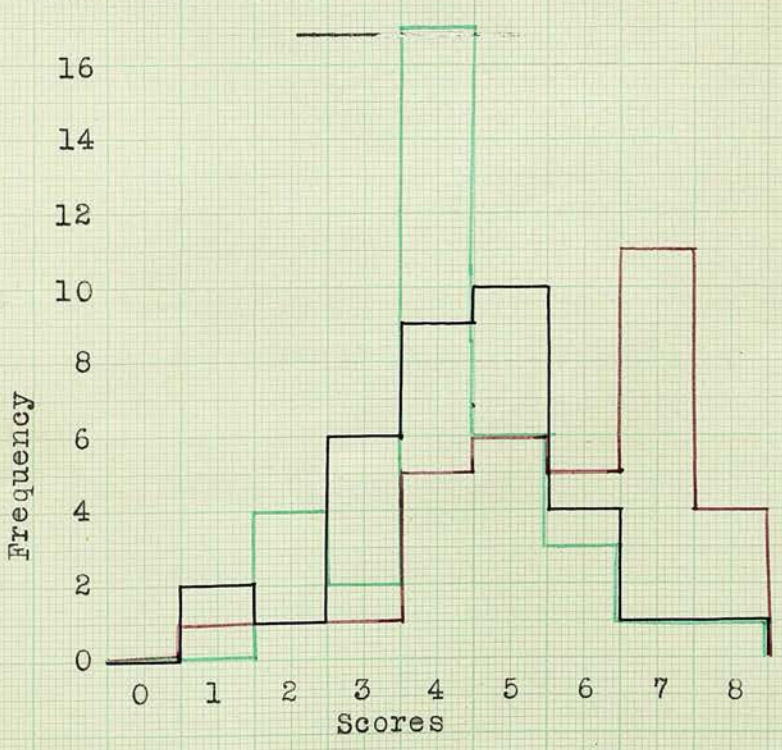
| | |
|---|---|
| Answer-pattern | 34, 33, 29, 27, 26, 24, 24, 19, 13. |
| Score-scatter | 0, 1, 1, 1, 5, 6, 5, 11, 4. |
| Standard deviation of A.P.D. | 2.66 |
| Standard deviation of score-scatter | 1.72 |
| Mean score | 5.7 |

It is obvious that the standard deviations of the scores vary as predicted. It is also interesting to note that the values fit quite well into the tail of the correlation table for the 10-item tests.

Fig. 16. Answer-patterns and score-scatters of Tests D, E, and F.




Answer-patterns.




Score-scatters.

## Chapter 7. The Correlation of the Coefficients of Skewness of Answer-pattern-differential and Score-scatter.

### 1. The measurement of skewness.

The skewness of a distribution such as a score-scatter is usually measured by the standardised value of the third moment about the mean. If the third moment $\frac{1}{n_0}\sum(x-a)^3$ is denoted by $\mu_3$, then the skewness S equals $\frac{\mu_3}{\sigma^3}$, where a, $n_0$, and $\sigma$ have their usual meanings.

The skewness so defined is positive when the curve has the tail to the right; negative when the curve has the tail to the left; and is zero for symmetrical distributions such as the normal frequency distribution.

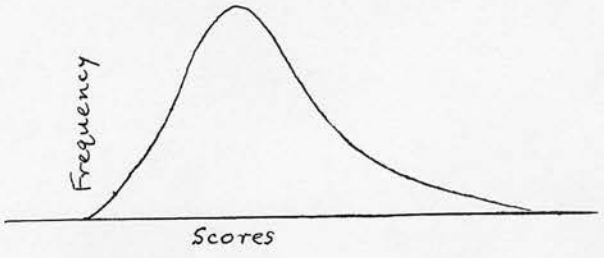



Figure 19. A positively skewed score-scatter.

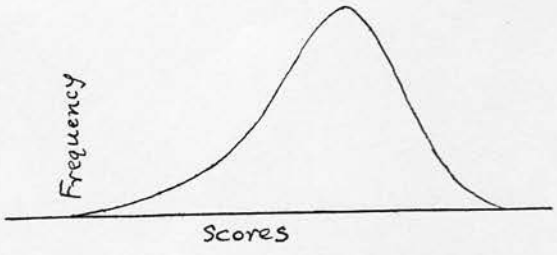



Figure 20. A negatively skewed score-scatter.

The application of this formula to the answer-pattern-differential may be justified in the same way as was done in the case of the standard deviations. The skewness of the answer-pattern-differential is denoted throughout this chapter by S'.

The method of calculation of S and S' follows closely that already shown for the standard deviations, except that the third moment must also be calculated. If the first, second, and third moments about the assumed mean are denoted by $m_1$, $m_2$, and $m_3$, then the skewness $\frac{\mu_3}{\sigma^3}$ can be shown to equal

$$\frac{m_3 - 3m_1m_2 + 2m_1^3}{(m_2 - m_1^2)^{3/2}}.$$

Thus there is constructed the following table, in which the data used are those of Test 4, which have already served in the previous chapter as an example.

Table 17. Calculation of S and S' for Test 4.

Answer-pattern-differential.

| Item | A.P. | A.P.D. | x | fx | $fx^2$ | $fx^3$ |
|---|---|---|---|---|---|---|
| 0 | 32 | 3 | -4 | -12 | 48 | -192 |
| 1 | 29 | 2 | -3 | - 6 | 18 | - 54 |
| 2 | 27 | 4 | -2 | - 8 | 16 | - 32 |
| 3 | 23 | 5 | -1 | - 5 | 5 | - 5 |
| 4 | 18 | 6 | 0 | 0 | 0 | 0 |
| 5 | 12 | 3 | 1 | 3 | 3 | 3 |
| 6 | 9 | 1 | 2 | 2 | 4 | 8 |
| 7 | 8 | 1 | 3 | 3 | 9 | 27 |
| 8 | 7 | 1 | 4 | 4 | 16 | 64 |
| 9 | 6 | 0 | 5 | 0 | 0 | 0 |
| 10 | 6 | 6 | 6 | 36 | 216 | 1296 |
| | 145 | $n_0$..32 | | +17 | +335 | +1115 |

$m_1 = 17/32 = +0.53$ ; $m_2 = 335/32 = +10.47$

$m_3 = 1115/32 = +34.84$

$$S' = \frac{m_3 - 3m_1m_2 + 2m_1^3}{(m_2 - m_1^2)^{3/2}} = +0.57$$

Score-scatter.

| Item | N | x | Nx | $Nx^2$ | $Nx^3$ |
|---|---|---|---|---|---|
| 0 | 1 | -4 | -4 | 16 | -64 |
| 1 | 1 | -3 | -3 | 9 | -27 |
| 2 | 1 | -2 | -2 | 4 | - 8 |
| 3 | 10 | -1 | -10 | 10 | -10 |
| 4 | 5 | 0 | 0 | 0 | 0 |
| 5 | 4 | 1 | 4 | 4 | 4 |
| 6 | 4 | 2 | 8 | 16 | 32 |
| 7 | 2 | 3 | 6 | 18 | 54 |
| 8 | 2 | 4 | 8 | 32 | 128 |
| 9 | 2 | 5 | 10 | 50 | 250 |
| 10 | 0 | 6 | 0 | 0 | 0 |
| | $n_0$..32 | | +17 | +159 | +359 |

$m_1 = +0.53$ $m_2 = +4.98$ $m_3 = +11.24$

$S = +0.36$

2. Results of the first experiment.

In this way the skewness of the answer-pattern-differential (S') and of the score-scatter (S) were calculated for each of the 41 tests. The results were as follows.

Table 18. Coefficients of skewness of the 41 tests.

| Test | S' | S |
|---|---|---|
| 1 | -0.015 | -0.28 |
| 2 | -0.09 | -0.17 |
| 3 | -0.38 | +0.15 |
| 4 | +0.57 | +0.36 |
| 5 | -0.52 | -0.003 |
| 6 | -0.19 | +0.33 |
| 7 | -0.09 | -0.01 |
| 8 | -0.76 | -0.32 |
| 9 | +0.24 | -0.06 |
| 10 | -0.603 | -0.42 |
| 11 | -1.08 | -0.78 |
| 12 | -0.64 | -1.12 |
| 13 | -0.42 | -0.23 |
| 14 | -0.35 | -0.59 |
| 15 | +0.14 | +0.103 |
| 16 | +0.206 | +0.02 |
| 17 | -1.002 | -0.55 |
| 18 | -0.23 | -0.14 |
| 19 | -0.41 | +0.04 |
| 20 | -0.05 | +0.21 |
| 21 | -1.01 | -0.297 |
| 22 | -0.15 | +0.12 |
| 23 | -0.92 | -0.93 |
| 24 | -0.69 | -0.28 |
| 25 | -0.93 | -0.74 |
| 26 | +0.99 | -0.74 |
| 27 | -0.799 | -0.27 |
| 28 | -0.59 | -0.202 |
| 29 | -0.52 | -0.78 |
| 30 | -0.55 | -0.23 |
| 31 | -1.06 | -0.95 |
| 32 | -0.95 | +0.07 |
| 33 | -1.31 | -0.38 |
| 34 | -0.75 | +0.004 |
| 35 | -0.53 | -0.13 |
| 36 | -0.53 | -0.35 |
| 37 | -0.55 | -0.28 |
| 38 | -0.43 | -0.13 |
| 39 | -0.42 | -0.404 |
| 40 | -0.94 | -0.64 |
| 41 | -0.48 | -0.69 |

The third decimal place has been added where required for grouping in the correlation table. This table is shown on page 91.

The correlation between S and S' was found to be $r = .628$. The probable error which might be attached to this coefficient is $\pm$ 0.064, or, using the transformation $z = \frac{1}{2}\log_e \frac{1+r}{1-r}$, we find that r differs significantly from .738 and .392. The correlation is not so high as that obtained with $\sigma_{A.P.D.}$ and $\sigma_N$ but it still is quite significant.

The regression equation is

$$S = 0.55S' + 0.00$$

and the standard deviation of an array of S is 0.28. The units used are those in which S and S' were originally measured.

3. The natural skewness of a population.

A rather interesting point may be dealt with here. Let us suppose that the population tested has a "natural" skewness, as opposed to the skewness forced on it by the answer-pattern. A possible way of measuring this natural skewness is to apply a test for which S' = 0; or, better still, to apply a battery of tests such as the 41 tests, and evaluate S from the regression equation obtained, substituting S' = 0 in that equation.

From the equation $S = 0.55S' + 0.00$ we find that the skewness of the population tested by the 41 tests is zero. This estimate has of course a standard deviation of 0.28, as seen above.

Table 19. Correlation of coefficients of skewness of answer-pattern-differential and score-scatter (41 tests). $r = .628$.

| S' \ S | -1.2 to -1.1 | -1.1 to -1.0 | -1.0 to -0.9 | -0.9 to -0.8 | -0.8 to -0.7 | -0.7 to -0.6 | -0.6 to -0.5 | -0.5 to -0.4 | -0.4 to -0.3 | -0.3 to -0.2 | -0.2 to -0.1 | -0.1 to -0.0 | 0.0 to 0.1 | 0.1 to 0.2 | 0.2 to 0.3 | 0.3 to 0.4 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 to 0.6 | | | | | | | | | | | | | | | | 1 | 1 |
| 0.4 to 0.5 | | | | | | | | | | | | | | | | | 0 |
| 0.3 to 0.4 | | | | | | | | | | | | | | | | | 0 |
| 0.2 to 0.3 | | | | | | | | | | | | 1 | 1 | | | | 2 |
| 0.1 to 0.2 | | | | | | | | | | | | | | 1 | | | 1 |
| 0.0 to 0.1 | | | | | | | | | | | | | | | | | 0 |
| -0.1 to 0.0 | | | | | | | | | | 1 | 1 | 1 | | | 1 | | 4 |
| -0.2 to -0.1 | | | | | | | | | | | | | | 1 | | 1 | 2 |
| -0.3 to -0.2 | | | | | | | | | | | 1 | | | | | | 1 |
| -0.4 to -0.3 | | | | | | | 1 | | | | | | | 1 | | | 2 |
| -0.5 to -0.4 | | | | | | 1 | | 1 | | 1 | 1 | | 1 | | | | 5 |
| -0.6 to -0.5 | | | | | 1 | | | | 1 | 3 | 1 | 1 | | | | | 7 |
| -0.7 to -0.6 | 1 | | | | | | | 1 | | 1 | | | | | | | 3 |
| -0.8 to -0.7 | | | | | | | | | 1 | 1 | | | 1 | | | | 3 |
| -0.9 to -0.8 | | | | | | | | | | | | | | | | | 0 |
| -1.0 to -0.9 | | | 1 | | 2 | 1 | | | | | | | 1 | | | | 5 |
| -1.1 to -1.0 | | | 1 | | 1 | | 1 | | | 1 | | | | | | | 4 |
| -1.2 to -1.1 | | | | | | | | | | | | | | | | | 0 |
| -1.3 to -1.2 | | | | | | | | | | | | | | | | | 0 |
| -1.4 to -1.3 | | | | | | | | | 1 | | | | | | | | 1 |
| Totals | 1 | 0 | 2 | 0 | 4 | 2 | 2 | 2 | 3 | 8 | 4 | 3 | 4 | 3 | 1 | 2 | 41 |

4. Second experiment.

We have reason to believe that mental ability like so many physical characteristics is distributed normally. Now the skewness of a normal distribution is zero. If the population tested is sufficiently large and unselected, it seems that we may assume its natural skewness to be zero, and any skewness found in the score-scatters to be attributable to the effect of a skewed answer-pattern-differential.

This method was applied to the complete tests already mentioned. The numbers tested are sufficiently large to obviate large deviations from the natural skewness zero, The fact that different candidates are under examination in each test does not affect the results.

In some cases the application of the method of calculation already shown was quite straightforward; in other cases, such as test M.H.T. 8, which has 109 items, the score-scatter and answer-pattern-differential had to be grouped in the usual way before evaluation of S and S'. There were 22 tests in all. The results are shown overleaf.

Table 20. Coefficients of skewness of complete tests.

| Test | S' | S |
|---|---|---|
| M.H.T. 8 | -0.38 | -0.66 |
| M.H.T. 9 | +0.27 | -0.06 |
| M.H.T. 11 | +0.15 | -0.15 |
| M.H.T. 12v | +0.199 | -0.17 |
| M.H.T. 12p | -1.201 | -1.33 |
| Thesis B | +0.23 | +0.01 |
| Thesis C | -0.28 | -0.24 |
| A II | -0.16 | -0.19 |
| A IX | -0.13 | -0.31 |
| A X | -0.36 | -0.32 |
| A XI | -0.28 | -0.38 |
| A XIII | -0.43 | -0.22 |
| A XV | +0.19 | -0.08 |
| A XXVI | -0.13 | -0.39 |
| A XXXIV | -0.26 | +0.07 |
| A XXXII | -0.12 | -0.15 |
| A XXXIII | +0.14 | -0.08 |
| K | +0.14 | +0.09 |
| L | -0.21 | +0.01 |
| K2 | -0.11 | -0.36 |
| M | -0.39 | -0.53 |
| M2 | -0.48 | -0.31 |

The correlation table showing the relation between S and S' is shown on page 94. The correlation was found to be r = .836 Proceeding as before we find that this coefficient, although derived from a comparatively small number of cases, is almost certainly greater than .627 and less than .929 .

The regression equation is

$$S = 0.77S' - 0.13 ,$$

with a standard deviation in the arrays of S of 0.17 .

Substituting S' = 0, we find that the natural skewness of the population tested is -0.13 , which does not differ significantly from zero, since the standard deviation of any array is 0.17 .

Table 21. Coefficients of skewness of answer-pattern-differential and score-scatter in correlation table form. (22 complete tests). $r = .836$.

| S' | S: -1.4 to -1.3 | -1.3 to -1.2 | -1.2 to -1.1 | -1.1 to -1.0 | -1.0 to -0.9 | -0.9 to -0.8 | -0.8 to -0.7 | -0.7 to -0.6 | -0.6 to -0.5 | -0.5 to -0.4 | -0.4 to -0.3 | -0.3 to -0.2 | -0.2 to -0.1 | -0.1 to 0.0 | 0.0 to 0.1 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 to 0.3 | | | | | | | | | | | | | | 1 | 1 | 2 |
| 0.1 0.2 | | | | | | | | | | | | | 2 | 2 | 1 | 5 |
| 0.0 0.1 | | | | | | | | | | | | | | | | 0 |
| -0.1 0.0 | | | | | | | | | | | | | | | | 0 |
| -0.2 -0.1 | | | | | | | | | | | 3 | | 2 | | | 5 |
| -0.3 -0.2 | | | | | | | | | | | 1 | 1 | | | 2 | 4 |
| -0.4 -0.3 | | | | | | | | 1 | 1 | | 1 | | | | | 3 |
| -0.5 -0.4 | | | | | | | | | | | 1 | 1 | | | | 2 |
| -0.6 -0.5 | | | | | | | | | | | | | | | | 0 |
| -0.7 -0.6 | | | | | | | | | | | | | | | | 0 |
| -0.8 -0.7 | | | | | | | | | | | | | | | | 0 |
| -0.9 -0.8 | | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | | | | | | | | | | | | | | 0 |
| -1.1 -1.0 | | | | | | | | | | | | | | | | 0 |
| -1.2 -1.1 | | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | 1 | | | | | | | | | | | | | | | 1 |
| -1.4 -1.3 | | | | | | | | | | | | | | | | |
| Totals | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 6 | 2 | 4 | 3 | 4 | 22 |

5. Significance of the experimental results.

It follows from the positive correlation between S and S' shown in the results of both these experiments that a positively skewed score-scatter is more likely to occur with a positively skewed answer-pattern-differential, and the extent to which it is skewed will depend partly on the degree of skewness of that distribution. To construct a test which is intended to give a score-scatter skewed positively, the examiner should therefore work with an answer-pattern falling steeply at first and then flattening out as it nears the items axis. Conversely, a test designed to produce a negatively skewed score-scatter should have an answer-pattern falling gently at first, and increasing in slope to a maximum. Then the nature of the correlation between S and S' found in the above experiments shows that the skewing of the answer-pattern-differential so caused will be accompanied by a similar skewing of the score-scatter to an extent indicated by the size of the correlation coefficient.

6. The influence of difficulty level on skewness of score-scatter

It is a fact already well known to examiners that score-scatters may be skewed by suitable adjustment of the difficulty level of a test. A recently published book on the science of marking says,

"If the curve is skewed to the low side, it means that marks have been difficult to get, which may be due.....to the questions

being too difficult. On the other hand, a curve skewed toward the upper part of the mark scale suggests that the paper has been easy. "

The problem that at once arises is the relation of this fact to the above theory. In the subsequent discussion I hope to make it clear that the use of the difficulty level as a method of skewing score-scatters is, in fact, merely an approximation to the use of the answer-pattern-differential.

The relation between difficulty level and skewness of score-scatter may be studied in the same way as the relation between S and S', provided we devise some measure of difficulty. This is comparatively easy. If the ratio of the total number of correct answers in a test to the total possible number is e, then the difficulty level d may be defined as 1 - e. In the notation used in previous chapters, $d = 1 - \frac{\sum n}{mn_o}$ .

The difficulties of all the tests used were calculated and are tabulated overleaf.

Table 22. Difficulty levels of tests.

(1) The 41 tests.

| Test | d | Test | d |
|---|---|---|---|
| 1 | .57 | 22 | .47 |
| 2 | .44 | 23 | .27 |
| 3 | .500 | 24 | .31 |
| 4 | .55 | 25 | .25 |
| 5 | .36 | 26 | .25 |
| 6 | .42 | 27 | .35 |
| 7 | .42 | 28 | .37 |
| 8 | .31 | 29 | .39 |
| 9 | .503 | 30 | .33 |
| 10 | .31 | 31 | .303 |
| 11 | .28 | 32 | .34 |
| 12 | .31 | 33 | .19 |
| 13 | .42 | 34 | .36 |
| 14 | .46 | 35 | .33 |
| 15 | .51 | 36 | .33 |
| 16 | .41 | 37 | .32 |
| 17 | .28 | 38 | .38 |
| 18 | .49 | 39 | .397 |
| 19 | .41 | 40 | .29 |
| 20 | .43 | 41 | .45 |
| 21 | .29 | | |

(2) The complete tests.

| Test | d | Test | d |
|---|---|---|---|
| M.H.T. 8 | .33 | A XIII | .44 |
| M.H.T. 9 | .56 | A XV | .56 |
| M.H.T. 11 | .505 | A XXVI | .51 |
| M.H.T. 12v | .52 | A XXXIV | .49 |
| M.H.T. 12p | .25 | A XXXII | .59 |
| Thesis B | .66 | A XXXIII | .58 |
| Thesis C | .53 | K | .52 |
| A II | .46 | L | .43 |
| A IX | .38 | K2 | .45 |
| A X | .39 | M | .42 |
| A XI | .41 | M2 | .46 |

Proceeding to the calculation of the correlation between the skewness (S) of the score-scatter and the difficulty level (d) of the answer-pattern, we find that for the 41 tests $r_{Sd}$ equals .561. and for the 22 tests equals .790 . These compare with $r_{SS'}$ = .628 for the 41 tests and .836 for the 22 tests. The correlation tables are shown on pages 99 and 100.

It will be observed that in each case the correlation between S and S' is greater than the correlation between S and d. On the other hand it must be pointed out that the difference is not statistically significant. To test the significance of the difference, it is necessary to use the transformation $z = \frac{1}{2}\log_e \frac{1+r}{1-r}$ For the 41 tests $r_{SS'}$ = .628, $r_{Sd}$ = .561; the corresponding values of z are .738 and .634, giving a difference .104. The standard deviation of this difference equals the square root of the sum of the squares of the standard deviations of the z's, and these deviations, as indicated in chapter 6, are each $\frac{1}{\sqrt{38}}$. The standard deviation of the difference is therefore $\sqrt{\frac{1}{38} + \frac{1}{38}} = .229$. Thus the difference is less than its standard error, and so is not significant. The lack of significance is even more strongly shown in the complete tests, their fewness increasing the standard deviation of the difference.

On the other hand, of all the various groups of tests which were afterwards selected from the whole 63 tests for various purposes, none has been found with $r_{SS'}$ less than $r_{Sd}$ .

Table 23. Correlation of S and d , 41 tests. $r = .561$.

| S | d: .18 to .21 | .21 to .24 | .24 to .27 | .27. to .30 | 30 to .33 | .33 to .36 | .36 to .39 | .39 to .42 | .42 to .45 | .45 to .48 | .48 to .51 | .51 to .54 | .54 to .57 | .57 to .60 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.3 to 0.4 | | | | | | | | | 1 | | | | 1 | | 2 |
| 0.2 0.3 | | | | | | | | | 1 | | | | | | 1 |
| 0.1 0.2 | | | | | | | | | | 1 | 1 | 1 | | | 3 |
| 0.0 0.1 | | | | | | 1 | 1 | 2 | | | | | | | 4 |
| -0.1 0.0 | | | | | | | 1 | | 1 | | 1 | | | | 3 |
| -0.2 -0.1 | | | | | | 1 | 1 | | 1 | | 1 | | | | 4 |
| -0.3 -0.2 | | | | 1 | 2 | 2 | 1 | | 1 | | | | | 1 | 8 |
| -0.4 -0.3 | 1 | | | | 1 | 1 | | | | | | | | | 3 |
| -0.5 -0.4 | | | | | 1 | | | 1 | | | | | | | 2 |
| -0.6 -0.5 | | | | 1 | | | | | | 1 | | | | | 2 |
| -0.7 -0.6 | | | | 1 | | | | | | 1 | | | | | 2 |
| -0.8 -0.7 | | | 2 | 1 | | | | 1 | | | | | | | 4 |
| -0.9 -0.8 | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | | 1 | 1 | | | | | | | | | | 2 |
| -1.1 -1.0 | | | | | | | | | | | | | | | 0 |
| -1.2 -1.1 | | | | | 1 | | | | | | | | | | 1 |
| Totals | 1 | 0 | 2 | 5 | 6 | 5 | 4 | 4 | 5 | 3 | 3 | 1 | 1 | 1 | 41 |

Table 24. Correlation of S and d, 22 tests. r = .790.

| S | d .24 to .27 | .27 to .30 | .30 to .33 | .33 to .36 | .36 to .39 | .39 to .42 | .42 to .45 | .45 to .48 | .48 to .51 | .51 to .54 | .54 to .57 | .57 to .60 | .60 to .63 | .63 to .66 | .66 to .69 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 to 0.1 | | | | | | | 1 | | 1 | 1 | | | | | 1 | 4 |
| -0.1 0.0 | | | | | | | | | | | 2 | 1 | | | | 3 |
| -0.2 -0.1 | | | | | | | | 1 | 1 | 1 | | 1 | | | | 4 |
| -0.3 -0.2 | | | | | | | 1 | | | 1 | | | | | | 2 |
| -0.4 -0.3 | | | | | 1 | 2 | 1 | 1 | 1 | | | | | | | 6 |
| -0.5 -0.4 | | | | | | | | | | | | | | | | 0 |
| -0.6 -0.5 | | | | | | | 1 | | | | | | | | | 1 |
| -0.7 -0.6 | | | | 1 | | | | | | | | | | | | 1 |
| -0.8 -0.7 | | | | | | | | | | | | | | | | 0 |
| -0.9 -0.8 | | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | | | | | | | | | | | | | | 0 |
| -1.1 -1.0 | | | | | | | | | | | | | | | | 0 |
| -1.2 -1.1 | | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | | | | | | | | | | | | | | | | 0 |
| -1.4 -1.3 | 1 | | | | | | | | | | | | | | | 1 |
| Totals | 1 | 0 | 0 | 1 | 1 | 2 | 4 | 2 | 3 | 3 | 2 | 2 | 0 | 0 | 1 | 22 |

7. The linearity of regression of S on various functions of d.

In the correlation of S with d , there is a difficulty which did not arise in previous correlations, the difficulty of linearity of regression. S and S' are both of the third order of moments, while d is of the first order. From this point of view it would seem that we should correlate S and $d^3$ instead of S and d. To avoid errors due to the introduction of non-linear regressions, all the correlations given above were tested for linearity, as also were the regression of S' on d, of S on $d^3$, and of S' on $d^3$ for both sets of data.

The method of testing the linearity of regression was that set out in Fisher's"Statistical Methods". This is a comparatively new method, replacing the older method using Blakeman's criterion, which was used in the author's second paper in the British Journal of Psychology. The objections to the validity of the older method are to be found in Fisher's book, section 46. In an appendix to this chapter there is shown in full the application of the new method to one of the above correlations.

For an understanding of the significance of the results it is necessary to give a very brief explanation of the method. Consider the correlation of S with d. For each value of d there is an array of values of S. The regression line of S on d is the straight line of best fit to the means of these arrays. The deviations of the means from this line represent the departures from linearity of regression; these deviations may be compared

with the deviations of the single observations from the mean of their array. Whether the deviations from linearity are significantly greater than those in the arrays is determined by calculating a variable z; then the probabilty of a given value of z being exceeded through chance variations is tabulated in Table VI of "Statistical Methods". The 5% point , representing a probabilty of 1 in 20 is usually taken as the dividing line.

When this test was applied to the various correlations, the results were in some ways surprising. The correlations $r_{SS'}$, $r_{Sd}$, $r_{S'd}$ for both groups of tests were sufficiently linear, with the exception of $r_{SS'}$ for the 22 tests. The value of z in this case was .70 while the 5% point was .53 and the 1% point .76. The deviations found would be exceeded only once in a hundred trials if the regression were linear. The lack of linearity is not due to any curving of the line of means; an examination of the correlation table shows it to be due rather to the zigzag nature of that line. It is probably caused by the mixing of data from various populations. An interesting point is that the correlation $r_{SS'}$ calculated from the 41 tests showed deviations from linearity less than the deviations within the arrays.

When the test was applied to the correlations involving $d^3$ it was found that all but one of these were also linear, the exception being the regression of S on $d^3$ for the 22 tests, where the value of z was .554 with a 5% point .548, so that the regressio is barely linear. Thus we have the apparent anomaly that S is

correlated in linear fashion with both d and $d^3$. This apparent anomaly will be cleared up later. Meanwhile, the test of linearity of regression applied to the data available does not enable us to distinguish between d and $d^3$ as to suitability for correlating with S.

8. The regression equation predicting S from S' and d.

Since S is found to be correlated both with S' and d, an obvious step is to construct the regression equation predicting S from S' and d. This equation is given by the formula

$$S = \frac{\Delta_{SS'}}{\Delta_{SS}} S' - \frac{\Delta_{Sd}}{\Delta_{SS}} d$$

where S, S', and d are measured in $\sigma$ units from their means and $\Delta_{SS'}$..... are the cofactors of $r_{SS'}$..... in the determinant

$$\Delta = \begin{vmatrix} r_{SS} & r_{SS'} & r_{Sd} \\ r_{SS'} & r_{S'S'} & r_{S'd} \\ r_{Sd} & r_{S'd} & r_{dd} \end{vmatrix}$$

The multiple correlation, R, of S with the team of S' and d weighted as above, is equal to $\sqrt{1 - \frac{\Delta}{\Delta_{SS}}}$

Before this equation can be set up, it is necessary to calculate the correlation of S' and d; this is readily obtainable from the data already given. From the 41 tests we find $r_{S'd}$ equal to .869 , and from the 22 tests $r_{S'd}$ equal to .805 . The tables are shown on pages 104 and 105.

Table 25. Correlation of S' and d, 41 tests.

| S' | d: .18 to .21 | .21 to .24 | .24 to .27 | .27 to .30 | .30 to .33 | .33 to .36 | .36 to .39 | .39 to .42 | .42 to .45 | .45 to .48 | .48 to .51 | .51 to .54 | .54 to .57 | .57 to .60 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 to 0.6 | | | | | | | | | | | | | 1 | | 1 |
| 0.4 0.5 | | | | | | | | | | | | | | | 0 |
| 0.3 0.4 | | | | | | | | | | | | | | | 0 |
| 0.2 0.3 | | | | | | | | 1 | | | 1 | | | | 2 |
| 0.1 0.2 | | | | | | | | | | | | 1 | | | 1 |
| 0.0 0.1 | | | | | | | | | | | | | | | 0 |
| -0.1 0.0 | | | | | | | | | 3 | | | | | 1 | 4 |
| -0.2 -0.1 | | | | | | | | | 1 | 1 | | | | | 2 |
| -0.3 -0.2 | | | | | | | | | | | 1 | | | | 1 |
| -0.4 -0.3 | | | | | | | | | | 1 | 1 | | | | 2 |
| -0.5 -0.4 | | | | | | | 1 | 2 | 1 | 1 | | | | | 5 |
| -0.6 -0.5 | | | | | 1 | 3 | 2 | 1 | | | | | | | 7 |
| -0.7 -0.6 | | | | | 3 | | | | | | | | | | 3 |
| -0.8 -0.7 | | | | | 1 | 1 | 1 | | | | | | | | 3 |
| -0.9 -0.8 | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | 2 | 2 | | 1 | | | | | | | | | 5 |
| -1.1 -1.0 | | | | 3 | 1 | | | | | | | | | | 4 |
| -1.2 -1.1 | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | | | | | | | | | | | | | | | 0 |
| -1.4 -1.3 | 1 | | | | | | | | | | | | | | 1 |
| Totals | 1 | 0 | 2 | 5 | 6 | 5 | 4 | 4 | 5 | 3 | 3 | 1 | 1 | 1 | 41 |

Table 26. Correlation of S' and d, 22 tests.

| S' | d .24 to .27 | .27 to .30 | .30 to .33 | .33 to .36 | .36 to .39 | .39 to .42 | .42 to .45 | .45 to .48 | .48 to .51 | .51 to .54 | .54 to .57 | .57 to .60 | .60 to .63 | .63 to .66 | .66 to .69 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 to 0.3 | | | | | | | | | | | 1 | | | | 1 | 2 |
| 0.1 0.2 | | | | | | | | | 1 | 2 | 1 | 1 | | | | 5 |
| 0.0 0.1 | | | | | | | | | | | | | | | | 0 |
| -0.1 0.0 | | | | | | | | | | | | | | | | 0 |
| -0.2 -0.1 | | | | | 1 | | 1 | 1 | 1 | | | 1 | | | | 5 |
| -0.3 -0.2 | | | | | | 1 | 1 | | 1 | 1 | | | | | | 4 |
| -0.4 -0.3 | | | | 1 | | 1 | 1 | | | | | | | | | 3 |
| -0.5 -0.4 | | | | | | | 1 | 1 | | | | | | | | 2 |
| -0.6 -0.5 | | | | | | | | | | | | | | | | 0 |
| -0.7 -0.6 | | | | | | | | | | | | | | | | 0 |
| -0.8 -0.7 | | | | | | | | | | | | | | | | 0 |
| -0.9 -0.8 | | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | | | | | | | | | | | | | | 0 |
| -1.1 -1.0 | | | | | | | | | | | | | | | | 0 |
| -1.2 -1.1 | | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | 1 | | | | | | | | | | | | | | | 1 |
| Totals | 1 | 0 | 0 | 1 | 1 | 2 | 4 | 2 | 3 | 3 | 2 | 2 | 0 | 0 | 1 | 22 |

The data from the 41 tests are thus

$r_{SS'} = .628$, $r_{Sd} = .561$, $r_{S'd} = .869$,

and the resulting equation is $S = 0.57\ S' + 0.06\ d$.

The multiple correlation R is .629 .

Similarly from the 22 tests with correlations

$r_{SS'} = .836$, $r_{Sd} = .790$, $r_{S'd} = .805$,

there is obtained the equation $S = 0.57\ S' + 0.33\ d$, and R equals .876.

The first regression equation states that S' is 0.57/0.06 times as important as d in predicting S, and that the correlation of S with the team of S' and d, weighted in the best possible way, is no greater than the correlation of S with S'.

In the case of the complete tests the result is not quite so definite. S' is about twice as important as d in predicting S, and the correlation of S with the team of S' and d is only slightly greater than the correlation of S with S' alone.

The most significant thing in the calculation is the high value of $r_{S'd}$. The large size of this correlation coefficient is the main cause of the team of S' and d being little superior to S' alone as a criterion of S. It seems that in measuring S' and d, we are, to a great extent, measuring the same thing. This, in fact, is the key to the whole problem.

9. The relation of S' and d.

The high correlation between S' and d might have been expected for the following reasons.

In an answer-pattern graph such as that shown below, the difficulty d of the test with answer-pattern (a) may be represented by the shaded area, for the area between the curve, OP, and OQ, is a measure of the total number of correct answers.

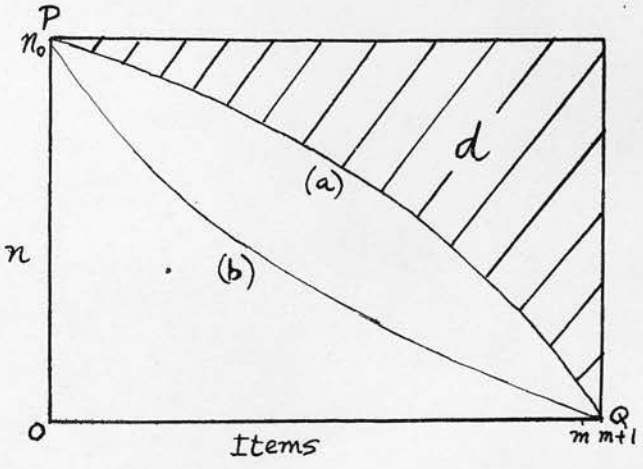



Figure 21. Answer-patterns of easy and difficult tests.

Now the answer-pattern of an easy test such as (a) will obviously produce a negatively skewed answer-pattern-differential On the other hand, a more difficult test, such as (b) will have a positively skewed answer-pattern-differential. It is possible to construct answer-patterns to which these arguments do not apply, but they are of peculiar type, such as tests half of whose items are answered correctly by all the candidates. These freak tests need not be considered here. In the ordinary test we may expect difficulty level and skewness of answer-pattern-differential

to be positively correlated.

In the case of a flat test, the skewness of the answer-pattern-differential is directly calculable from the difficulty level. Consider a test of m items, sat by $n_o$ candidates, each item being answered correctly n times. The answer-pattern is then $n_o$, n, n, .....n, n. The calculation of the skewness proceeds as before (cf. page 88)

| Item (x) | A.P. | A.P.D. | fx | $fx^2$ | $fx^3$ |
|---|---|---|---|---|---|
| 0 | $n_o$ | $n_o-n$ | 0 | 0 | 0 |
| 1 | n | 0 | 0 | 0 | 0 |
| 2 | n | 0 | 0 | 0 | 0 |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| m-1 | n | 0 | 0 | 0 | 0 |
| m | n | n | mn | $m^2n$ | $m^3n$ |
| | | $n_o$ | mn | $m^2n$ | $m^3n$ |

The moments are $m_1 = \frac{mn}{n_o}$

$$m_2 = \frac{m^2n}{n_o}$$

$$m_3 = \frac{m^3n}{n_o}$$

Substituting these values in the formula

$$S' = \frac{m_3 - 3m_1m_2 + 2m_1^3}{(m_2 - m_1^2)^{3/2}}$$

and using the identities

$$d = 1 - \frac{mn}{mn_o} = 1 - \frac{n}{n_o}$$

we find

$$S' = \frac{2d-1}{\sqrt{d(1-d)}}$$

This function of d may be denoted by $\phi(d)$. For any flat test the skewness of the answer-pattern-differential is therefore directly calculable from the difficulty level. Consider, for example, a test in which each item is answered correctly by half the candidates. Here $n = n_o/2$ and therefore $d = 0.5$. Substituting in the formula we find $S' = 0$. The answer-pattern-differential of such a test is therefore unskewed.

In passing, we may note that the formula suggests that the function of d which should be correlated with S is neither d nor $d^3$ but $\phi(d)$. To check this point, $\phi(d)$ was evaluated for the tests under consideration, and the linearity of the correlations $r_{s\phi}$ was tested. From the 41 tests the value of $r_{s\phi}$ obtained was .576 ( cf. $r_{Sd} = .561$, $r_{Sd^3} = .490$ ), and the correlation was certainly linear ( $z = .15$, 5% point .38 ). From the 22 tests the value of $r_{s\phi}$ obtained was .816 ( cf. $r_{Sd} = .790$, $r_{Sd^3} = .620$ ) and the correlation was again linear, ( $z = .34$, 5% point .55 ). The increase in the size of the correlation coefficient points to the employment of a better statistical method.

Where the test is not perfectly flat, the skewness of the answer-pattern-differential is not equal to $\phi(d)$. But it can be demonstrated that in any test $\phi$ is a good approximation to S', and d is a good approximation to $\phi$. The close relation of S' and d is a corollary of the theorem that d is an approximation, through $\phi$, to S'.

It will readily be seen that the first approximation, of $\phi$ for $S'$, consists in replacing the answer-pattern of the test, from which S' is calculable, by the answer-pattern of a flat test of the same difficulty level. Instead of using all the distinctive values of n which make up the answer-pattern, the tester uses only the average of these values. Symbolically, the process is as follows.

The skewness S' of the answer-pattern-differential is given by the formula

$$n_0\sigma^3 S' = \sum_{x=0}^{m}(n_x - n_{x+1})(x - a)^3$$

where a is the mean score.

Let n be the mean value of $n_1, n_2, \ldots\ldots\ldots, n_m$,

and define $\nu_x$ as $n_x - n$.

Then $$n_x - n_{x+1} = \nu_x - \nu_{x+1}$$

and $$n_m = \nu_m + n$$

$$n_0\sigma^3 S' = (\nu_0 - \nu_1)(0-a)^3 + (\nu_1 - \nu_2)(1-a)^3 + \ldots\ldots\ldots$$
$$\ldots + (\nu_{m-1} - \nu_m)(m-1-a)^3 + (\nu_m + n)(m-a)^3.$$
$$= -\nu_0 a^3 + \nu_1(3a^2-3a+1) + \ldots + \nu_m\left[3(m-a)^2-3(m-a)+1\right] + n(m-a)^3.$$

Now $\nu_0 > \nu_1 \geqslant \nu_2 \geqslant \nu_3 \geqslant \ldots$ as far as $\nu_x$ is positive,

and $a^3 > 3a^2-3a+1 > \ldots\ldots\ldots$ since $a > 1$.

Therefore the first term is greater than the second, and so on.

Also $n > |\nu_m| \geqslant |\nu_{m-1}| \geqslant \ldots\ldots\ldots$ where $\nu_x$ is negative,

and $(m-a)^3 \geqslant 3(m-a)^2-3(m-a)+1 \geqslant \ldots$ since $m > a+1$.

Therefore the last term is greater than the second last, and so on.

A first approximation to S' is therefore given by using only the first and last terms of the series.

$$\therefore n_o \sigma^3 S' = -\gamma_o a^3 + n(m-a)^3.$$

Dividing through by $n_o m^3$, and using the identities

$$d = 1 - \frac{a}{m} = 1 - \frac{n}{n_o},$$

we have $$\frac{\sigma^3}{m^3} S' = -d(1-d)^3 + (1-d)d^3$$

i.e. $$S' = \frac{m^3}{\sigma^3} \left[ d(1-d)(2d-1) \right]$$

This is precisely the formula previously obtained for $\phi(d)$, once the substitution for $\sigma$ has been made.

The reliability of this approximation may be tested by calculating $r_{S'\phi}$ from the available data. The 41 tests gave r = .867, and the 22 tests r = .817 .

As already indicated, the process of approximation may be carried a stage further. In the expression for $\phi(d)$, put $d = \frac{1}{2} + \delta$ . Then

$$\phi(d) = \frac{2d-1}{\sqrt{d(1-d)}} = \frac{4\delta}{\sqrt{1-4\delta^2}} = 4\delta \left( 1 + 2\delta^2 + \ldots\ldots\ldots \right).$$

Now $|\delta| < \frac{1}{2}$ ; as a first approximation we may therefore take

$$\phi(d) = 4\delta = 4d - 2,$$

which is a linear function of d.

The error introduced by this latter approximation varies with the size of $\delta$ , that is with the value of d. For d equal to 0.3 or 0.7, i.e., $\delta = \pm 0.2$, the error is about 10%, and it decreases as $|\delta|$ decreases.

Within these limits, $\phi$ is practically a linear function of d. The great majority of the tests under consideration had difficulty levels within this range. This is probably the reason for the apparent anomaly of S being correlated in linear fashion with such varying functions of d as $d^3$, $\phi(d)$, and d itself.

An interesting commentary on the above approximation is provided by the regression equation predicting S' from d, as calculated from the data of the 41 tests. When this equation is put into the form where S' and d are measured in their original units it becomes $S' = 4.10\ d - 2.07$ .

This long and rather involved discussion therefore leads to the conclusion that the difficulty level d is a first approximation to $\phi(d)$, which in turn is a first approximation to S', the skewness of the answer-pattern-differential. The correlation found to exist between S, the skewness of the score-scatter, and d, the difficulty level, is an effect of the correlation existing between S and S'.

## 10. Conclusion.

The wellknown rule-of-thumb method of skewing score-scatters by the variation of difficulty level is an approximation to the basic principle, that the skewing of score-scatters is controlled by the skewness of the answer-pattern-differential. Since it is an approximation, this method cannot be expected to yield such accurate results as that using the answer-pattern-differential; this is demonstrated by the diminution in the

correlations shown in the table below.

Table 27. Correlation of S with S', $\phi$, and d.

| | 41 tests | 22 tests |
|---|---|---|
| $r_{SS'}$ | .628 | .836 |
| $r_{S\phi}$ | .576 | .816 |
| $r_{Sd}$ | .561 | .790 |

For this reason a better prediction of S will always be obtained from S' than from d alone. From a study of the whole answer-pattern one gains a better idea of the probable skewness of the score-scatter than would be obtained from a knowledge of the test's difficulty alone.

Appendix. Fisher'S Method of testing Linearity of Regression.

Since this method is of fairly recent origin, an example is worked out in full below, the regression tested being that of S on d for the 41 tests. The correlation table is on page 99.

For each value of d there is an array of values of S. Let us designate any array by the suffix p; the number of observations in the array may be denoted by $n_p$, and the mean of the array by $\bar{S}_p$; the mean of all the values of S is denoted by $\bar{S}$. Then it may be shown that

$$\sum( S - \bar{S} )^2 = \sum\sum( S - \bar{S}_p )^2 + \sum\{n_p ( \bar{S}_p - \bar{S} )^2\}.$$

This is an algebraic identity, expressing the fact that the total variation of S may be split up into two parts, one representing the variation of the observations about the mean

of their array; the other the variation of the means of the arrays about the general mean. The variation of the means about the general mean is partly due to the slope of the regression line, and the amount of this variation is calculable; the other part is due to deviations from linearity. For each variation the standard deviation is calculable, and the test of significance of the deviations from linearity becomes the test of whether the standard deviation of the deviations from linearity is, or is not, significantly greater than the standard deviation of the observations within the array about their mean.

In the correlation table on page 99, number the arrays from left to right, omitting the array in which there are no entries. For ease of calculation, an assumed mean may be taken at S = -0.35, and the class intervals are taken as unit. Then for each array of S, the excess is calculated, and this is also done for the total. The results of these calculations are shown in the table overleaf.

The sum of the mean squared excess less the mean square of the total excess gives $\sum n_p (\bar{S}_p - \bar{S})^2$ , which here equals 300.8 .

The total variation of S has already been found in the calculations of the correlation; it is 532.9 . Thus the variation $\sum\sum( S - \bar{S}_p )^2$ equals 532.9 - 300.8 = 232.1 .

Table 28. Calculation of $\sum n_p ( \bar{S}_p - \bar{S} )^2$.

Arrays of S for given values of d.

| S | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -8 | | | | 1 | | | | | | | | | | 1 |
| -7 | | | | | | | | | | | | | | 0 |
| -6 | | | 1 | 1 | | | | | | | | | | 2 |
| -5 | | | | | | | | | | | | | | 0 |
| -4 | | 2 | 1 | | | | 1 | | | | | | | 4 |
| -3 | | | 1 | | | | | | 1 | | | | | 2 |
| -2 | | | 1 | | | | | | 1 | | | | | 2 |
| -1 | | | | 1 | | | 1 | | | | | | | 2 |
| 0 | 1 | | | 1 | 1 | | | | | | | | | 3 |
| 1 | | | 1 | 2 | 2 | 1 | | 1 | | | | | 1 | 8 |
| 2 | | | | | 1 | 1 | | 1 | | 1 | | | | 4 |
| 3 | | | | | | 1 | | 1 | | 1 | | | | 3 |
| 4 | | | | | 1 | 1 | 2 | | | | | | | 4 |
| 5 | | | | | | | | | 1 | 1 | 1 | | | 3 |
| 6 | | | | | | | | 1 | | | | | | 1 |
| 7 | | | | | | | | 1 | | | | 1 | | 2 |
| $n_p$ | 1 | 2 | 5 | 6 | 5 | 4 | 4 | 5 | 3 | 3 | 1 | 1 | 1 | 41 |
| Excess | 0 | -8 | -14 | -13 | +8 | +10 | +3 | +19 | 0 | +10 | +5 | +7 | +1 | +28 |
| $\frac{(\text{Excess})^2}{n_p}$ | 32 | 392 | 282 | 128 | 25 | | 22 | 722 | 0 | 333 | 25 | 49 | 1 | 19.1 |

Total = 319.9 - 19.1 = 300.8

For each of these terms there must also be found the number of degrees of freedom. The number of observations is 41, therefore the number of degrees of freedom for the total variation is 40. The number of arrays is 13, therefore the number of degrees of freedom for the variation between arrays is 12. By subtraction the number of degrees of freedom within the arrays must be 28 .

Of the variation between the arrays part is due to the slope of the regression line. This may easily be shown to be equal to $r^2 \sum ( S - \bar{S} )^2$, which in the present case is 167.8 . The number

of degrees of freedom represented here is 1.

The variation between the arrays due to deviations from linearity is therefore equal to 300.8b- 167.8 = 133.0, and the corresponding number of degrees of freedom is 12 - 1 = 11 . We have therefore to decide whether a variation of 133.0 obtained from 11 degrees of freedom is significantly greater than a variation of 232.1 obtained from 28 degrees of freedom.

This is the same problem as determining whether an estimate of standard deviation derived from $n_1$ degrees of freedom is significantly greater than a second estimate obtained from $n_2$ degrees of freedom. The method is to evaluate z equal to the difference of the natural logarithms of the two standard deviations, i.e., $z = \log_e \frac{\sigma_1}{\sigma_2}$ ; then the probability of exceeding this value of z by chance is tabulated in Fisher's Table VI for given values of $n_1$ and $n_2$ . As before, a probability of .05 is taken as the dividing line.

In the present example the calculation may be completed thus

| Variance of S | | Degrees of freedom | Mean square | $\frac{1}{2}\log_e$ |
|---|---|---|---|---|
| Total | 532.9 | 40 | | |
| Between arrays | 300.8 | 12 | | |
| ∴ Within arrays | 232.1 | 28 | 8.29 | 1.06 |
| Due to linear regression | 167.8 | 1 | | |
| ∴ Due to deviations | 133.0 | 11 | 12.09 | 1.24 |
| | | | | z = 0.18 |

For $n_1$ = 11, $n_2$ = 28, the 5% point is 0.39; that is, the probability of exceeding z = 0.39 by chance is 1 in 20. The regression of S on d, with z = 0.18 is definitely linear.

## Chapter 8. The Measurement of the Control of Score-scatter by Answer-pattern in single tests.

### 1. The nature of the problem.

In chapter 2 it was shown that the answer-pattern-differential and the score-scatter of a test were identical in the particular case referred to as "unig", where each candidate's score was compiled of answers to the easiest possible questions. The results submitted in chapters 5, 6, and 7 show that in the more general case, where randomness of answering is present, there is still some measure of relation or correspondence between the two distributions, though there is no longer identity. The measure of correspondence in each case was the correlation coefficient of corresponding statistics of the distribution. The calculation of these correlations implies the existence of several, and if possible of many, tests given to the same or similar populations.

The problem to be studied in this chapter is the measurement of the degree of correspondence of the two distributions in a single test. Given the results of a single test, is it possible to estimate the degree of control the answer-pattern has exerted on the score-scatter ?

First it may be pointed out that the answer-pattern-differential and score-scatter are so removed from identity that no use can be made of Pearson's test of Goodness of Fit. In practically every case this test when applied to the answer-pattern-differential

and score-scatter of a test, would indicate these to be entirely different distributions. In any case, Pearson's test can only be used to decide whether or not two distributions may be regarded as identical save for the effects of sampling, and cannot be used to measure the degree of control or goodness of fit of the two, as the name might suggest.

It is necessary, then, to devise some other method of measuring the relationship that undoubtedly exists even in the case of a single test. The problem is so beset with difficulties that, although four coefficients have been devised, evaluated for all the tests, and otherwise used, none appears satisfactory. Only two will be mentioned here.

2. The coefficient of hig -"h"

This coefficient was devised and used by the author in studies for the degree of Bachelor of Education.

When there is exact correspondence between answer-pattern-differential and score-scatter, the equations (B) of chapter 2 hold. That is,

$n_x - n_{x+1} = N_x$ holds for x = 0, 1, 2,.....m.

The extent of deviations from this state may then be measured by the expression

$$\sum_{x=0}^{m} ( n_x - n_{x+1} - N_x )^2$$

which obviously vanishes when equations (B) hold.

Conversely, since the expression is the sum of squares, when it vanishes each term must vanish; i.e.

$n_x - n_{x+1} = N_x$ is true for x = 0, 1, 2, .....m,

that is there is exact correspondence between answer-pattern-differential and score-scatter. The equations also imply that the test is unig, i.e. that each candidate's score must be made up of answers to the easiest questions. For, the last equation is $n_m = N_m$, i.e. those, and only those, making the maximum score have answered the last question. Since $n_{m-1} - n_m = N_{m-1}$ or $n_{m-1} = n_m + N_{m-1}$, the penultimate question must have been answered by $n_m + N_{m-1}$ candidates. Now these include the $N_m$ candidates who made perfect scores, and also the $N_{m-1}$ who scored m-1, since these latter could not have answered question m as an alternative. Therefore the number of times question m-1 is answered is accounted for entirely by those candidates scoring m-1 and upwards. A similar argument extends to scores m-2 and so on. The equation $\sum ( n_x - n_{x+1} - N_x )^2 = 0$ therefore implies the unig type of answering.

The expression $\sum ( n_x - n_{x+1} - N_x )^2$ has a minimum value of zero when the scores are unig. It is at a maximum when the scores are made in random or hig fashion. The probability of this, as was proved in chapter 2, is greatest when the answer-pattern is flat, i.e. when $n_1 = n_2 = n_3 = ..... = n_m = n$ say.

Then the sum becomes

$$\sum ( n_x - n_{x+1} - N_x )^2 = \sum_0^m N_x^2 + n^2 + (n_o - n)^2 - 2nN_m - 2(n_o - n)N_o$$

To obtain a coefficient which will make it possible to compare tests with differing numbers of candidates and differing

numbers of items, let us define the coefficient of hig as

$$h = \frac{\sum(n_x - n_{x+1} - N_x)^2}{\sum N_x^2 + n^2 + (n_o - n)^2 - 2nN_m - 2(n_o - n)N_o}$$

where n is the mean of $n_1$, $n_2$, .....$n_m$.

This coefficient is independent of the number of candidates sitting the test; for if the candidates be augmented by a similar population the values of n and N will be increased in the same ratio, and the expression for h being homogeneous in n and N remains the same. The effect of varying the number of items is rather complicated and will not be investigated here. In actual results from tests it has been found that the coefficient calculated directly from the data of a 100 item test does not differ much from the coefficient obtained from the same data after grouping the scores and answer-pattern-differentials into 10 groups of 10, i.e., replacing the 100 item test by a 10 item test.

The method of calculation may be illustrated from the data of Test 4 of the 41 tests, already used as an example in previous chapters.

Table 29. Calculation of coefficient of hig h .

| x | $n_x$ | $n_x - n_{x+1}$ | $N_x$ | $N_x^2$ | $\lvert n_x - n_{x+1} - N_x \rvert$ | $\lvert\lvert^2$ |
|---|---|---|---|---|---|---|
| 0 | 32 | 3 | 1 | 1 | 2 | 4 |
| 1 | 29 | 2 | 1 | 1 | 1 | 1 |
| 2 | 27 | 4 | 1 | 1 | 3 | 9 |
| 3 | 23 | 5 | 10 | 100 | 5 | 25 |
| 4 | 18 | 6 | 5 | 25 | 1 | 1 |
| 5 | 12 | 3 | 4 | 16 | 1 | 1 |
| 6 | 9 | 1 | 4 | 16 | 3 | 9 |
| 7 | 8 | 1 | 2 | 4 | 1 | 1 |
| 8 | 7 | 1 | 2 | 4 | 1 | 1 |
| 9 | 6 | 0 | 2 | 4 | 2 | 4 |
| 10 | 6 | 6 | 0 | 0 | 6 | 36 |
| | 145 | | | 172 | | 92 |

$n = 145/10 = 14.5$ $\qquad \Sigma N^2 = 172$

$n_o - n = 17.5$ $\qquad n^2 = 210$

$(n_o - n)^2 = 306$

$688$

$N_o = 1$ $\qquad 2(n_o - n)N_o = 35$

$653$

$N_m = 0$

$h = 92/653 = 0.14$

---

The values of h in the 41 tests ranged from .06 to .41, the median being .24 . In the complete tests h ranged from .06 to .40, the values for some of these tests were

| | |
|---|---|
| M.H.T. 8 | .25 |
| M.H.T. 9 | .11 |
| M.H.T. 11 | .10 |
| M.H.T. 12v | .16 |
| M.H.T. 12p | .20 |
| Thesis A | .40 |
| Thesis B | .09 |
| Thesis C | .07 |

3. Criticism of h.

There are two weak points in the definition of h. First, in the process of obtaining the denominator representing the incidence of maximum hig, the answer-pattern of the test was altered to a flat answer-pattern of the same difficulty level. The score-scatter was left unchanged. Now the evidence of chapter 4 indicates that the answer-pattern of a test is just as permanent a feature as the score-scatter; in altering the answer-pattern we have changed the whole test.

At bottom, this weakness seems to depend on a confusion of ideas between hig as poorness of fit of answer-pattern-differential and score-scatter, and hig as randomness of answering due to lack of agreement between the individual examinee's order of difficulty of items and the average order of difficulty. These two ideas are of course linked by the unig case where there is perfect fit and perfect agreement.

The second weak point is as fundamental. It has already been noted that there is a natural tendency of score-scatters to normality. An answer-pattern-differential which is already of this shape will apparently have a better chance of showing a good fit to any score-scatter obtained than would an answer-pattern-differential of a completely different shape, say the answer-pattern-differential of a flat test. Thus the low values of h obtained with tests M.H.T. 9, M.H.T. 11, Thesis B and C may be attributed to the shapes of their answer-patterns, giving

answer-pattern-differentials already akin to normal distributions. On the other hand the high value of h obtained from test A may be attributed to the shape of A's answer-pattern-differential, which was of the type shown in figure 22.

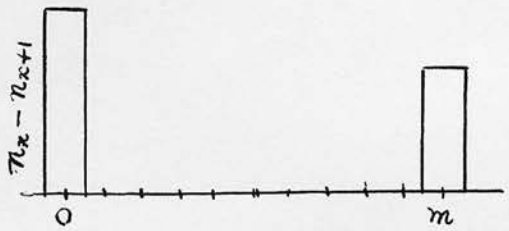



Figure 22. Answer-pattern-differential of a flat test.

Although h is large in the case of test A, it may be that the answer-pattern of the test actually exerted more control in the production of the score-scatter than did the answer-patterns of tests B and C, but the effect is masked by tendency of score-scatters to normality.

4. Definition of $\alpha$.

In the formation of the score-scatter of a test there seem then to be two factors, the answer-pattern-differential of the test, and what might be called the "natural" distribution of the scores. This is the distribution that would be obtained if the influence of the answer-pattern could be removed. From what we have learned we should postulate it as a normal distribution, which implies that it is unskewed; and its mean is of course fixed as equal to the mean of the answer-pattern-differential, or what is the same thing, the mean of the score-scatter formed. Though the mean and skewness of the distribution are so fixed,

there is no method of estimating its standard deviation, even when the tests are mental tests and the standard deviation of intelligence quotients is taken as 13 points.

Let us suppose meantime that this distribution is known; that is for each score x there is known the value of $w_x$, the number of candidates making that score. Then the actual score-scatter obtained with anyvtest may be regarded in its most simple form as a linear function of the answer-pattern-differential and the natural score-scatter; that is

$$N_x = \alpha (n_x - n_{x+1}) + \beta w_x$$

where $\alpha, \beta$ are parameters. The ratio $\alpha/\beta$ is a measure of the relative influence of answer-pattern-differential and natural score-scatter in the formation of N.

In a ten item test there are eleven such equations, the values of $\alpha$ and $\beta$ varying with the equation. These may be regarded as eleven equations for $\alpha$ and $\beta$, and from them have to be derived single values of $\alpha$ and $\beta$ which best represent the position for the whole test. The number of equations is too small to enable us to set up a regression equation determining the best values of $\alpha$ and $\beta$; in any case, the variables $N_x$, $n_x - n_{x+1}$ and $w_x$ are not normally distributed.

The best values of $\alpha$ and $\beta$ may be determined by the method of least squares. Each equation is multiplied by the coefficient of $\alpha$, and the resulting equations are added to give one equation in $\alpha$ and $\beta$. Similarly, by multiplying each

equation by the coefficient of $\beta$ and adding, a second equation in $\alpha$ and $\beta$ is obtained. The solution of these two simultaneous equations gives the values of $\alpha$ and $\beta$ which fit best all the original equations. If $n_x - n_{x+1}$ is denoted for brevity by $v_x$ the equations may be written

$$\alpha = \frac{\sum Nv \sum w^2 - \sum Nw \sum vw}{\sum v^2 \sum w^2 - (\sum vw)^2}$$

$$\beta = \frac{\sum Nw \sum v^2 - \sum Nv \sum vw}{\sum v^2 \sum w^2 - (\sum vw)^2}$$

In the application of this formula to the 41 tests the first difficulty is the construction of the w distributions appropriate to each test. As noted before, the mean and skewness of these distributions is fixed, but some value for the standard deviation must be assumed before the distribution can be calculated. Since no theoretical basis has so far been found on which the standard deviation could be estimated, some rule of thumb method must be used meantime. The following method was that finally chosen for the 41 tests. It was found that the average standard deviation of the scores in these tests was 2.12 . This also approximates closely to the standard deviation of scores that would most probably have been obtained from a test of that type, the items of which progressed uniformly in difficulty from very easy to very hard. In such a test it is easy to show that the standard deviation of the answer-pattern-differential is 3.16 , and the corresponding standard deviation of scores obtained from the regression equation for the 41 tests is 2.10. A value near 2.10

had therefore both a theoretical and practical significance for the 41 tests. To ease the calculation of w somewhat, the standard deviation finally chosen was 2.00 .

Thereafter the method of calculation followed closely that demonstrated in the appendix to chapter 2. As a sample, there is shown below the table worked out for test 4. The scores were grouped in twos to decrease random errors due to the smallness of the frequencies in some of the classes.

The mean score, obtained from the answer-pattern-differential before grouping, was 4.5 .

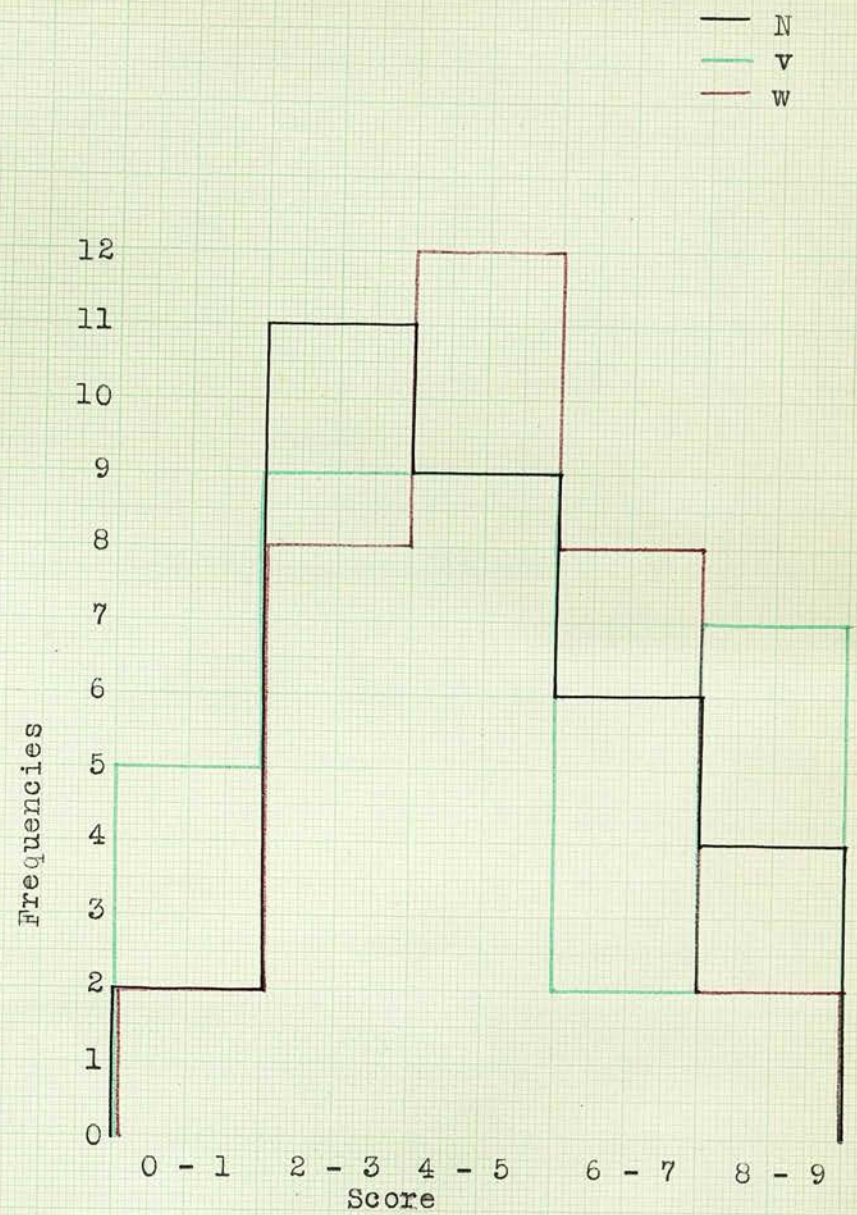
Table 30. Calculation of w distribution for test 4.

| Score | x | $\frac{x-a}{\sigma}$ | I | Difference | 32 × Difference |
|---|---|---|---|---|---|
| | | $[-\infty]$ | -.500 | | |
| 0-1 | | | | .067 | 2 |
| | 1.5 | -1.5 | -.433 | | |
| 2-3 | | | | .242 | 8 |
| | 3.5 | -0.5 | -.191 | | |
| 4-5 | | | | .382 | 12 |
| | 5.5 | +0.5 | +.191 | | |
| 6-7 | | | | .242 | 8 |
| | 7.5 | +1.5 | +.433 | | |
| 8-9 | | | | .067 | 2 |
| | | $[+\infty]$ | +.500 | | |
| | | | | | 32 |

For test 4 the normal distribution was then taken as above, 2,8,12,8,2 for the appropriate scores.

Now the use of an answer-pattern-differential 5,9,9,2,7 altered the above normal score-scatter to that actually observed i.e., 2,11,9,6,4. These distributions are shown in histogram form overleaf ( page 127 ).

Fig. 21. Distributions N, v, and w for Test 4.



The calculation of $\alpha$ and $\beta$ then proceeds as follows.

Table 31. Calculation of $\alpha$ and $\beta$ for test 4.

| Score | N | v | w | $v^2$ | $w^2$ | Nv | Nw | vw |
|---|---|---|---|---|---|---|---|---|
| 0-1 | 2 | 5 | 2 | 25 | 4 | 10 | 4 | 10 |
| 2-3 | 11 | 9 | 8 | 81 | 64 | 99 | 88 | 72 |
| 4-5 | 9 | 9 | 12 | 81 | 144 | 81 | 108 | 108 |
| 6-7 | 6 | 2 | 8 | 4 | 64 | 12 | 48 | 16 |
| 8-9 | 4 | 7 | 2 | 49 | 4 | 28 | 8 | 14 |
| Totals | 32 | 32 | 32 | 240 | 280 | 230 | 256 | 220 |

$\alpha = 0.43$, $\beta = 0.58$

In test 4 then, the relative influences of answer-pattern-differential and of tendency to normality are in the ratio 43/58.

As it is found that $\alpha + \beta$ approximates closely to unity in every case, the value of $\alpha$ alone may be given as a sufficient indication. It was found in the 41 tests that the value of $\alpha$ ranged from -0.22 (test 24) to +0.50 (test 39) with a median value +0.14 .

The weaknesses of this method of determining the relative strengths of answer-pattern-differential and normal tendency are obvious. It demands a precise knowledge of the "natural" distribution which we do not have at the present stage. In its calculation there arise difficulties such as the calculation of w for tests with high or low means, causing lack of headroom or the opposite. The theoretical basis of the natural distribution is flimsy. In referring to a score-scatter uninfluenced by any answer-pattern-differential we are creating a mathematical fiction.

5. Conclusion.

These difficulties bring us to consider the necessity of constructing such coefficients as h and $\alpha$. Their usefulness is based on the following train of reasoning. There has been proved to be a certain measure of agreement between the answer-pattern-differential of a test and its score-scatter. The examiner who can select his test items from a battery of items of known difficulty can therefore predict within certain limits the nature of the score-scatter that will be obtained when the test is given to a known population. It may be that tests vary in the extent of the agreement between their answer-pattern-differential and score-scatter ( though that is by no means definitely proved yet), some tests showing closer agreement, and others a greater tendency for the score-scatter to vary. Then the examiner would tend to choose that type of test for which the agreement was good so that his attempt to procure a given score-scatter would be more likely to be successful. In defining h, $\alpha$ and other coefficients not mentioned here, we are seeking some way of labelling tests so that, having divided the sheep from the goats, we may analyse the factors causing some to be sheep and the others goats. It may be, as stated above, that there is no such distinction; that the fluctuations in agreement are due merely to errors of sampling, and the small numbers of tests that have perforce had to be used render this possibility quite feasible. More light is shed on the problem by the results of the next chapter.

## Chapter 9. The Relation of Steepness of Tests to the Control of Score-scatter by Answer-pattern.

### 1. The definition of steepness.

The problem of what type of test gives the best agreement of answer-pattern-differential and score-scatter may be tackled in a different way from that adopted in the last chapter. It was shown in chapter 2 that the probability of unig was greatest when the test was of the steep type, i.e. one in which $n_1$ is much greater than $n_2$, which in turn is much greater than $n_3$, and so on. Since unig implies perfect fit of score-scatter and answer-pattern-differential it would seem a priori that the steeper tests should show closer agreement of score-scatter and answer-pattern-differential than would be obtained with flatter tests.

The definition of "steep" is sofar no more than the bare statement that in such a test $n_1$ is much greater than $n_2$, $n_2$ than $n_3$, and so on. What this implies may conveniently be considered in two steps.

(1) It is obvious that steepness depends to a great extent on the number of items in the test. A test with few items has, a priori, a much greater chance of being made steep than a test with many items, since in the latter case it is impossible to make great differences of difficulty between adjacent items. Consider, for example, the following three item test.

Simplify (1) $2 + 3 - 6$

(2) $\frac{4}{3}\left\{\frac{5}{8} + \frac{1}{2}\right\}$

(3) $\frac{ab}{a+b}\left\{\frac{a}{b} + \frac{b}{a}\right\}$

This is a steep test. It is almost certain that the answers to such a test would be of unig type; any candidates answering correctly question 3 would answer correctly questions 1 and 2, and so on. The introduction of additional items of intermediate difficulty would obviously lessen the degree of certainty, and if the number of items were increased to ten, even though these made use of the full range of difficulty between item 1 and item 3 of the above test, there would obviously be little hope of the answers being completely unig.

This state of matters is reflected in the probability of hig, as defined on page 32. For a test of three items, with an answer-pattern $n_0 = 100$, $n_1 = 80$, $n_2 = 50$, $n_3 = 20$, the probability of unig is 0.6 . For a very similar test with the same general shape of answer-pattern but having four items, $n_1 = 80$, $n_2 = 60$, $n_3 = 40$, $n_4 = 20$, the probability of unig is only 0.2 . For a ten-item test of any shape of answer-pattern, the probability of unig is so low as to be negligible.

(2) Although the number of items to be used may be so large that the probability of unig is negligible, it may be that tests with steeper answer-patterns,(i.e. with difficulty differences as large as possible in view of the number of items present )

show closer agreement between the answer-pattern-differential and the score-scatter. These tests will be referred to as "steeper" rather than as "steep". For a study of this it will be necessary to define more exactly what is meant by steepness in the case of tests with equal numbers of items.

Consider the answer-patterns illustrated in the diagram.

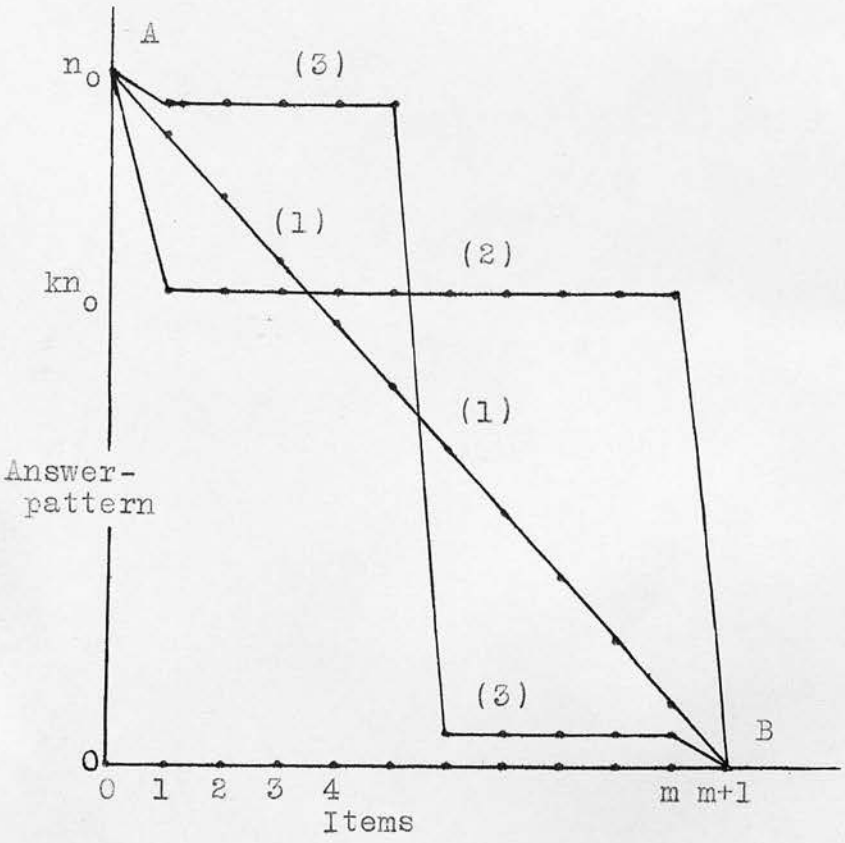



Figure 24. Answer-patterns of various types of test.

(1) is the answer-pattern of what might be called the steepest possible test for the given number of items. The points lie on a straight line joining $x=0$, $y=n_o$ to $x=m+1$, $y=0$. Strictly the curve making the probability of unig a maximum is not a straight line, but the whole calculation is of such mathematical

difficulty, and the result so far as obtained so nearly a straight line, that this has been adopted.

(2) is the answer-pattern of a flat test, each item of which has been answered by $km_0$ of the $n_0$ candidates. $(k<1)$

(3) represents an answer-pattern partly steep and partly flat.

It will be obvious from these diagrams that it is impossible to measure the steepness of a test merely by averaging the difference in difficulty between adjacent items. Such an average would equal $\frac{n_1 - n_m}{m - 1}$ for all the tests shown, and would rank tests 1 and 3 as of equal steepness.

After several methods of defining steepness had been tried and abandoned, the following mode of definition was devised. The points of the answer-pattern of the steepest test (1) lie on a straight line AB. The length of the path from A to B through these answer-pattern points is therefore a minimum. In the case of the other answer-patterns, the length of the path from A to B is obtained by adding the lengths of the straight lines joining adjacent points of the answer-pattern, including the line from A and that to B. The length of this path is an index of the flatness of the test.

In test (1), the length of the path, l, equals $\sqrt{n_0^2 + (m+1)^2}$ in test (2) it equals $\sqrt{(1-k)^2 n_0^2 + 1} + (m-1) + \sqrt{k^2 n_0^2 + 1}$ which is approximately equal to $(1-k)n_0 + m - 1 + km_0 = n_0 + m - 1$ independent of the value of k. This may be taken as the

maximum length of the path, but for our purpose it is even better to take $n_0 + m + 1$ as the maximum length. The advantage of this choice is that it makes easier the construction of a coefficient which will be independent of the number of candidates sitting the test. For the same reason it is necessary to fix the scale of the diagram by taking as a representative number of candidates $m + 1$. This makes the triangle AOB isosceles and the minimum length of the path becomes $(m+1)\sqrt{2}$, and the maximum length $2(m+1)$.

Suppose that the original answer-pattern of any test has the values $n_0, n_1, n_2, \ldots\ldots n_m$, and denote the corresponding answer-pattern-differentials $n_x - n_{x+1}$ by $\Delta_x$. Then, after adjustment of the diagram scale to make the triangle isosceles, the differences will be $\frac{m+1}{n_0}\Delta_x$, and the length of the path from A to B will be $\sum_{x=0}^{m}\sqrt{1 + \left(\frac{m+1}{n_0}\Delta_x\right)^2}$.

The coefficient of steepness $c$ may now be defined as the ratio*

$$c = \frac{\text{maximum length - actual length}}{\text{maximum length - minimum length}}$$

$$= \frac{2(m+1) - \sum\sqrt{1 + \left(\frac{m+1}{n_0}\Delta_x\right)^2}}{2(m+1) - \sqrt{2}(m+1)}$$

$$= \frac{1}{.293}\left\{1 - \frac{1}{2(m+1)}\sum\sqrt{1 + \left(\frac{m+1}{n_0}\Delta_x\right)^2}\right\}$$

The calculation of this coefficient for any of the 41 tests is a very simple matter. It is shown for test 11 overleaf.

*So defined to make $c = 0$ for a flat test, and $c = 1$ for the steepest test.

Table 32. Calculation of c for test 11.

| Answer-pattern | A.P.D. $\Delta$ | $\frac{11}{32}\Delta$ | $\sqrt{1+\left(\frac{11}{32}\Delta\right)^2}$ |
|---|---|---|---|
| 32 | 5 | 1.72 | 1.99 |
| 27 | 1 | .34 | 1.08 |
| 26 | 0 | 0 | 1.00 |
| 26 | 2 | .69 | 1.21 |
| 24 | 0 | 0 | 1.00 |
| 24 | 1 | .34 | 1.08 |
| 23 | 0 | 0 | 1.00 |
| 23 | 0 | 0 | 1.00 |
| 23 | 1 | .34 | 1.08 |
| 22 | 9 | 3.10 | 3.26 |
| 13 | 13 | 4.47 | 4.58 |
| | | | 18.28 |

$$c = \frac{1}{.293}\left(1 - \frac{18.28}{22}\right) = 0.6$$

This coefficient was calculated for each of the 41 tests, and the values obtained ranged from 0.6 (test 11) to 0.9 ( test 20 ).

3. The influence of steepness on the fit of answer-pattern-differential and score-scatter.

By the use of this measure of steepness, the 41 tests may be divided into two groups of steeper and flatter tests, the dividing line being fixed so as to place approximately equal numbers of tests in each group. Each of the groups then yields data from which can be calculated the correlation of the standard deviations of the answer-pattern-differential and the score-scatter (as was done in chapter 6 for the 41 tests), and the correlation of the coefficients of skewness of these distributions ( as was done in chapter 7). When these were

calculated the results were as follows.

The correlation of the standard deviations of answer-pattern-differential and of score-scatter for the 20 steeper tests was r = .776; for the 21 flatter tests it was r = .830. This difference is not significant, the difference of the corresponding values of z being .15 and its standard deviation being .34 . In the matter of standard deviations the flatter tests thus show a fit which is better , but not significantly so, than do the steeper tests.

The correlation of the skewness of the answer-pattern-differential with the skewness of the score-scatter for the 20 steeper tests was r = .795; for the 21 flatter tests it was r = .590. This difference, though greater than that found with the standard deviations, is still not significant: the difference in the values of z is .41, and its standard deviation is .34 . Here the better fit is obtained from the steeper tests.

From the scanty data at our disposal it would seem then that the control of score-scatter by answer-pattern-differential is no greater in the case of the steeper group than in the case of the flatter group. This conclusion is strengthened by the result obtained in a second method of examining the position in these 41 tests.

The coefficient $\alpha$ defined in chapter 8, measured the relative influence of the answer-pattern-differential and a hypothetical normal distribution on the score-scatter obtained. By correlating this coefficient with the coefficient c we may

ascertain whether the steeper tests show a relatively greater influence of answer-pattern-differential. This correlation was calculated from the data of the 41 tests, and was found to be $r = +.218$, which though positive is not significantly different from zero, being derived from 41 pairs of values.

It would appear then that for tests of ten items or more, the control of score-scatter by answer-pattern-differential is independent of the steepness of the test. In a way, this is rather a welcome conclusion, since it simplifies matters somewhat As will be shown in the next chapter, it is necessary to use flat answer-patterns to produce certain types of score-scatter, If the certainty of control decreased with the flatness of the answer-patterns the position would be much more confused than it is. As far as these results go, they show that a flat answer-pattern exerts about as much control over the score-scatter as does any other type.

On the other hand the results show that to be really steep in the sense of producing answers approaching unig type, a test must have very few items, say three or four, and these spaced out in the best way. This is a type of test few examiners would care to use; it would very easily give rise to difficulties through misunderstandings, prior knowledge, and guesswork, all of which may be smoothed out in a test with many items.

We are thus in the position of having "gained upon the roundabouts what we lost upon the swings". We have lost hope of constructing useful unig tests, but have gained the knowledge that

flat tests show as great a correspondence between answer-pattern and score-scatter as do the steeper tests.

4. An advantage in the use of steeper tests.

There is one advantage of the steeper test which is worth mentioning. It sometimes happens that a test designed for a group of given average ability is employed to test a group of slightly different ability. It is easy to show by an example that in such a case the steeper test has its characteristics changed less by the altered character of the testees than has a flat test of the same average difficulty.

Consider two tests which when applied to a certain group of candidates yield the following answer-patterns.

Test 1

| Item | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 100 | 91 | 82 | 73 | 64 | 55 | 46 | 37 | 28 | 19 | 10 |

This represents the steepest possible test of ten items. Tha answer-pattern-differential has an average score 5, a standard deviation 3.16 and skewness 0.00.

Test 2

| Item | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 100 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |

This represents a flat test, with an average score 5, a standard deviation of answer-pattern-differential 5.00 and skewness again 0.00 .

Suppose now that these two tests are given to a group of 100 candidates, whose mean ability is less than that of the original group by an amount sufficient to depress the percentage of correct answers to the items of test 2 from 50 to 40. This corresponds, in the case of an intelligence test, to an age difference of four months at age eleven. Using the technique described in chapter 4 we can now calculate the changes in the answer-patterns of both tests.

The answer-pattern of test 1 becomes

100, 86, 75, 65, 55, 45, 37, 28, 20, 13, 5,

yielding a mean score 4.29, and an answer-pattern-differential with standard deviation 3.12 and skewness +0.30.

The answer-pattern of test 2 becomes

100, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40.,

yielding a mean score 4.00, and an answer-pattern-differential with standard deviation 4.90 and skewness +0.41 .

It will readily be seen that in each respect the steeper test has suffered less change than the flat one. This superiority of the steeper test in permanence of answer-pattern will tend to be reproduced in the score-scatters, since these are in part controlled by the answer-patterns.

Another feature of the steeper tests is that their reliability as measured by the split-halves method is likely to be greater This question will be discussed in a later chapter.

## Chapter 10. The Construction of Answer-patterns.

### 1. The theoretical basis.

From the preceding theory it is apparent that the examiner who wishes to produce a score-scatter of a given type must construct an answer-pattern appropriate to his purpose. This assumes that he has access to a battery of items of known difficulty for the population considered: such a collection has been made, for instance, by Professor Thorndike and is described in his "Measurement of Intelligence", and a similar collection might be compiled from the data of the Moray House series of Intelligence tests, English tests, and Arithmetic tests.

Since the main body of data used in the present investigation comprises tests of ten items each we shall use as an example a ten item test. Suppose that an examiner wishes to construct a ten item test producing a score-scatter specified by its mean, its standard deviation, and its skewness. The mean of the corresponding answer-pattern-differential is then fixed as equal to the mean of the intended score-scatter. The standard deviation and the skewness of the answer-pattern-differential to be used must be calculated from the regression equations of chapters 6 and 7, or, more easily, read off from the regression lines.

There are thus three quantities given, sufficient to fix only three points of the answer-pattern. The first stage must then be the construction of a three point answer-pattern, such as would be obtained from a three item test, with a mean three-tenths of the

given mean, a standard deviation diminished in the same ratio, but an unaltered skewness, since that is already measured in standardised units.

Let the ordinates of this three item answer-pattern be $n_0$, $n_1$, $n_2$, $n_3$; the answer-pattern -differential is therefore $n_0-n_1$, $n_1-n_2$, $n_2-n_3$, $n_3$. Let $y_0 = \frac{n_0 - n_1}{n_0}$, $y_1 = \frac{n_1 - n_2}{n_0}$, and so on. Then

$$y_0 + y_1 + y_2 + y_3 = 1 .$$

If the first, second, and third moments of this y distribution about zero are denoted by $m_1$, $m_2$, and $m_3$, then

$$\begin{aligned} y_1 + 2y_2 + 3y_3 &= m_1 \\ y_1 + 4y_2 + 9y_3 &= m_2 \\ y_1 + 8y_2 + 27y_3 &= m_3 \end{aligned}$$

Now it is easy to prove that $m_1 = a$, $m_2 = \sigma^2 + a^2$, $m_3 = \sigma^3 S + 3\sigma^2 a + a^3$, where $a$, $\sigma$, and $S$ are the mean, standard deviation, and skewness of the three item answer-pattern-differential.

Thus we obtain the four equations

$$\begin{aligned} y_0 + y_0 + y_2 + y_3 &= 1 \\ y_1 + 2y_2 + 3y_3 &= m_1 = a \\ y_1 + 4y_2 + 9y_3 &= m_2 = \sigma^2 + a^2 \\ y_1 + 8y_2 + 27y_3 &= m_3 = \sigma^3 S + 3\sigma^2 a + a^3 . \end{aligned}$$

From these equations the values of $y_0$, $y_1$, $y_2$, $y_3$ may be calculated. The solution is made easier if the m's are

first calculated from the given values of a, $\sigma$, and S. Then the required values of the y's interms of the known m's are;

$$y_0 = 1 - \tfrac{11}{6}m_1 + m_2 - \tfrac{1}{6}m_3$$
$$y_1 = 3m_1 - \tfrac{5}{2}m_2 + \tfrac{1}{2}m_3$$
$$y_2 = -\tfrac{3}{2}m_1 + 2m_2 - \tfrac{1}{2}m_3$$
$$y_3 = \tfrac{1}{3}m_1 - \tfrac{1}{2}m_2 + \tfrac{1}{6}m_3$$

From these values of y the values of n giving the required answer-pattern are easily calculated for any given number of candidates. An example will make the application of the method clearer, and will serve as a basis for the discussion of the processes involved in converting this three item pattern into a ten item pattern.

2. An example.

It is desired to construct a ten item answer-pattern giving an answer-pattern-differential with mean score 4, standard deviation 4, and skewness +0.5 .

The corresponding values for the three item answer-pattern-differential are a = 1.2, $\sigma$ = 1.2, S = +0.5 .

Hence $m_1 = 1.2$, $m_2 = 2.88$, $m_3 = 7.77$ .

Hence $y_0 = 0.38$, $y_1 = 0.29$, $y_2 = 0.07$, $y_3 = 0.26$.

For 100 candidates this would mean an answer-pattern-differential 38, 29, 7, 26; that is, an answer-pattern 100, 62, 33, 26.

It is profitable for us to analyse a little more fully some of the steps in the above calculation, to bring out their

implications. When the mean was fixed at 1.2, $m_1$ was thereby fixed at 1.2, and the values of $m_2$ and $m_3$ were also partially fixed, since they are functions of the mean and other variables. Similarly fixing the standard deviation at 1.2 finally fixed the value of $m_2$, and still further circumscribed the range of possible values of $m_3$, or in other words delimited the range of possible values of S. The limitation of the range of possible values of S arises through each y being a positive (or zero) quantity not more than unity.

If, in the example, we substitute the values for a and $\sigma$, and then evaluate $y_0$, $y_1$, $y_2$, $y_3$, we can find the limits within which S must lie. The equations become

$$y_0 = .525 - .288S$$
$$y_1 = -.144 + .864S$$
$$y_2 = .504 - .864S$$
$$y_3 = .112 + .288S$$

The limits of each y are 0 and 1, by the definition of answer-patterns. Hence we deduce from the respective equations the following pairs of rough limits of S.

$+2 > S > -2 \; : \; +0.2 < S < +1 \; : \; +0.6 > S > -0.6 \; : \; -0.4 < S < +3$ .

Taken together the four equations limit the permissible values of S to the range +0.2 to +0.6. The value actually chosen was +0.5 .

The process might have been reversed. S might have been first fixed, then $\sigma$, and finally the mean a would have had to be

chosen from a restricted range. The three variables concerned may be inserted in any order.

The next step is the conversion of this three item pattern to a ten item pattern. In this process there is lacking that exactness of method and fixity of results which characterized the preceding part of the calculation. There are many possible ways of filling out a three item answer-pattern to one of ten items. Of these, there are two that merit further investigation. They form the extremes, as regards steepness, of the possible methods.

The first is to give each of the first three items of the ten item test the same difficulty as the first item of the three item test; items four to seven the difficulty of the second item of the three item test; and items eight to ten the difficulty of the last item of the small test. In the example considered, the ten item pattern would be

100, 62, 62, 62, 33, 33, 33, 33, 26, 26, 26.

This gives an answer-pattern-differential

38, 0, 0, 29, 0, 0, 0, 7, 0, 0, 26,

with a mean 3.96, standard deviation 4.05, and skewness +0,55. The steepness of this test as measured by the coefficient c is 0.5 . It is the flattest test that can be constructed to fulfil the required conditions.

A second method of filling out the answer-pattern is to use the given values as representative points on a smooth curve, and

from this curve determine the values of the n's required. The representative points are placed at x = 2, $5\frac{1}{2}$, 9. The curve for the test considered is on page 146. From it we derive the answer-pattern

100, 78, 62, 51, 43, 36, 31, 29, 27, 26, 25,

which gives an answer-pattern-differential

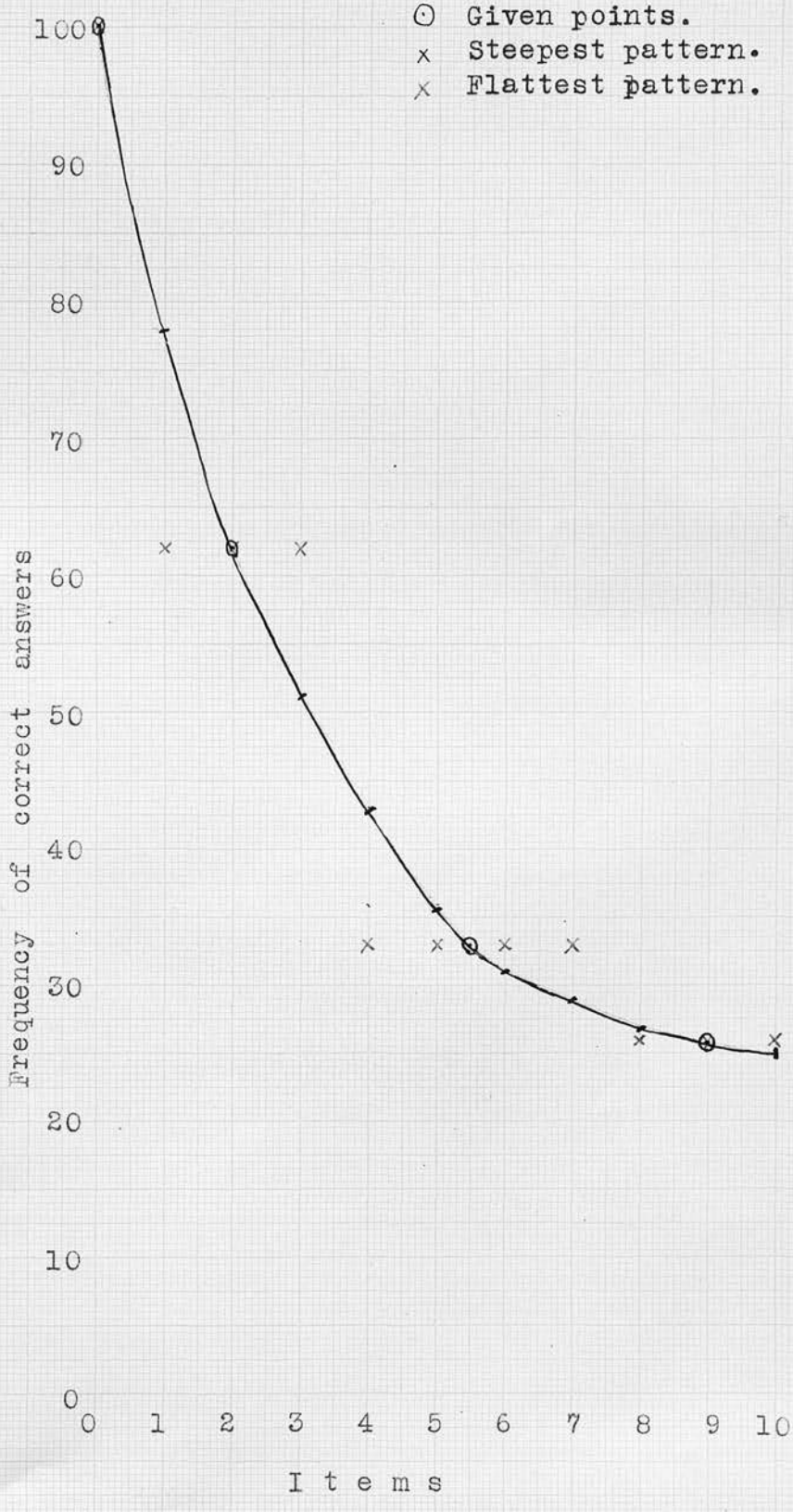
22, 16, 11, 8, 7, 5, 3, 1, 1, 1, 25,

with mean 4.07, standard deviation 3.89, and skewness +0.40. This is the steepest answer-pattern that can be constructed to fulfil the required conditions; its coefficient of steepness c equals 0.8 .

The tests which might be constructed to have the required mean, standard deviation, and skewness, are therefore restricted in steepness to the range 0.5 to 0.8 . As far as we know, there is no reason to prefer the steeper or the flatter type of test, save that a preference might be shown for the steeper test on the grounds mentioned on page 139. The final choice of the examiner would probably be determined by the items available.

If a test of 100 items were required, it would be necessary to amke use of regression equations connecting standard deviation of answer-pattern-differential and that of score-scatter, and connecting the coefficients of skewness of these distributions. It has been shown in chapter 7 that the equation for skewness is practically the same for 100 item tests as for 10 item tests. In the case of standard deviations this is not the case, probably

Fig. 25. Completion of answer-pattern from given points.



because of the lack of "headroom" in the 10 item tests. It would be necessary first to collect sufficient data to construct a reliable regression equation for 100 item tests. This the author has been unable to do yet.

## Chapter 11. The Relation of the Reliability of Tests to the Nature of the Answer-pattern.

### 1. Theoretical relations of reliability and hig.

The reliability of tests of the type under consideration is usually measured by the split-halves method. If the correlation between the scores on the odd items and those on the even items is denoted by $r_{\frac{1}{2}}$, then the coefficient of reliability is defined as $\frac{2r_{\frac{1}{2}}}{1 + r_{\frac{1}{2}}}$. It represents the correlation that would exist between the scores made in the complete test and those made in a similar test if such existed.

There are obviously some points of connection between this coefficient and the answer-pattern of the test. It is interesting to note that the present investigation arose from a discussion among certain educationists, one of whom claimed that the reliability of tests such as group tests of intelligence was a fictitious reliability, its magnitude being largely preordained by the tester eliminating as far as possible differences due to hig. The author was asked to investigate this problem, which soon resolved itself into a much wider series of problems, and the fruits of the research are submitted in this thesis. The original problem may now appear to be rather a bypath from the main track, but it is quite important from the point of view of those constructing and using tests of the type discussed here.

Before the reliability coefficient is calculated, the items

of the test should be placed in order of difficulty. Under these conditions it is possible to prove that unig is sufficient but not necessary, to produce perfect reliability.

(1) Unig implies perfect reliability.

If the answers to the test are unig each score 2x is made up of answers to the x easiest odd items and the x easiest even items. Thus there is perfect correlation between scores on odd and even items. Scores which are odd, and therefore cannot be halved exactly, will produce small variations which may in practice depress the correlation slightly below the theoretical maximum.

(2) Unig is not necessary for perfect reliability.

This is a negative proposition which is most easily proved by constructing a particular case, a test which has perfect reliability and yet some degree of hig. Such a test may be constructed as under.

Consider a 6 item test, the items being numbered 1,3,5,7,9,11 Suppose that the following results are obtained when the test is attampted by 10 candidates A - J.

(Table 33)

Table 33. Results of a 6 item test.

| Candidates | Items 1 | 3 | 5 | 7 | 9 | 11 | Scores |
|---|---|---|---|---|---|---|---|
| A | | | x | | | | 1 |
| B | | x | | x | | | 2 |
| C | x | | | | x | | 2 |
| D | x | | x | | x | | 3 |
| E | x | x | x | | | | 3 |
| F | x | x | x | | | | 3 |
| G | x | x | x | | | | 3 |
| H | x | x | x | x | | | 4 |
| I | x | x | | x | | x | 4 |
| J | x | x | | x | x | x | 5 |
| | 8 | 7 | 6 | 4 | 3 | 2 | 30 |

Answer-pattern

| Score-scatter | x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|
| | $N_x$ | 0 | 1 | 2 | 4 | 2 | 1 | 0 |

It is apparent from the table that a certain amount of hig is present; h, as defined in chapter 8, equals .28 .

This test could be converted into a 12 item test of perfect reliability merely by doubling the table, making item 2 a duplicate of item 1 and so on. This would not be a satisfactory method of proof, as the resulting test would be of a most unusual type. In any case such a procedure is quite unnecessary, as it is quite easy to construct another test with the same answer-pattern

and score-scatter. Such a test is shown below, the items being numbered 2,4,6,8,10,12. Again h = .28 .

Table 34. Results of a 6 item test.

| Candidates | Items 2 | 4 | 6 | 8 | 10 | 12 | Scores |
|---|---|---|---|---|---|---|---|
| A | x | | | | | | 1 |
| B | | | x | | x | | 2 |
| C | x | x | | | | | 2 |
| D | x | x | | | | x | 3 |
| E | x | x | x | | | | 3 |
| F | x | | x | x | | | 3 |
| G | x | x | x | | | | 3 |
| H | x | x | | x | x | | 4 |
| I | x | x | x | x | | | 4 |
| J | | x | x | x | x | x | 5 |
| Answer-pattern | 8 | 7 | 6 | 4 | 3 | 2 | 30 |

| Score-scatter | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| $N_x$ | 0 | 1 | 2 | 4 | 2 | 1 | 0 |

If the results of these two tests are now combined to give a 12 item test, the results are as shown in the following table. It will be observed that each candidate's score may be equally divided into answers to odd and even items. That is, the reliability of the test is perfect.

Table 35. Combined results giving a 12 item test.

| Candidates | Items 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Scores |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | x | | | x | | | | | | | | 2 |
| B | | | x | | | x | x | | | x | | | 4 |
| C | x | x | | x | | | | | x | | | | 4 |
| D | x | x | | x | x | | | | x | | | x | 6 |
| E | x | x | x | x | x | x | | | | | | | 6 |
| F | x | x | x | | x | x | | x | | | | | 6 |
| G | x | x | x | x | x | x | | | | | | | 6 |
| H | x | x | x | x | x | | x | x | | x | | | 8 |
| I | x | x | x | x | | x | x | x | | | x | | 8 |
| J | x | | x | x | | x | x | x | x | x | x | x | 10 |
| | 8 | 8 | 7 | 7 | 6 | 6 | 4 | 4 | 3 | 3 | 2 | 2 | 60 |

Answer-pattern

| Score-scatter | x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N_x$ | | 0 | 0 | 1 | 0 | 2 | 0 | 4 | 0 | 2 | 0 | 1 | 0 | 0 |

It might be objected that this is a very artificially constructed test, and that it is most unlikely that any test could be so neatly split up into the two components here shown. The reply to this objection is that it is merely another way of saying that a test with perfect reliability is unlikely to occur at all. What has been proved is that, if it did occur, there is no necessity for the test to be unig. It seems that the test is at least just as likely to show hig as to be unig.

When the items of the test are not in order of difficulty, unig is neither necessary nor sufficient to produce perfect

reliability. That unig is not necessary may be proved directly by suitable rearrangement of the items of the above test, care being taken to keep the odd items as odd, or to change all the odd items to even items, and vice versa. There will then be formed a test of perfect reliability which has a certain amount of hig in the answers.

A general argument may be used to establish the insufficiency of unig for perfect reliability when the items are not in order of difficulty. When the order of items is random, it is unlikely that any score of 2x, though made up of answers to the 2x easiest items, should be made up of answers to x odd and x even items, or that any more complicated relation between scores on odd and even items should exist causing perfect correlation. There are m! - 1 ways of rearranging the m items of a unig test. Of these $\left\{\left(\frac{m}{2}\right)!\right\}^2 - 1$ preserve odds as odds and evens as evens. The probability of the reliability being still perfect after rearrangement is therefore $\left(\left\{\left(\frac{m}{2}\right)!\right\}^2 - 1\right)/(m! - 1)$, which for m= 10 roughly equals 1 in 252.

The sole positive conclusion that has been drawn, then, is that unig type of answering implies perfect reliability when the items are in order of difficulty. This may mean that tests in which the incidence of hig has been reduced will have on that account a higher reliability coefficient, but such a conclusion must be established by evidence from experiments.

2. Experimental evidence on the relation of reliability and hig.

An initial difficulty here is that no really satisfactory measure of the quantity of hig in a test has yet been devised. There have been described the coefficient of hig, h; the coefficient $\alpha$ measuring the ratio of the influence of the answer-pattern-differential to that of a hypothetical normal distribution; and the coefficient of steepness, c, which may bear some relation to the amount of hig present. These must serve meantime.

(a) Data of 41 tests.

In the case of the 41 tests all these coefficients are available. The reliability of one of those tests must be calculated by placing the items in order of difficulty, so classifying the items as odd or even, and then correlating the scores on odd and even items. These scores must, of course, be obtained from the original data.

Each of the 41 tests is a 10 item test, so that the scores to be correlated are those on the 5 odd items with those on the 5 even items. This naturally leads to a coarseness of grouping effect in the correlation table, with an adverse effect on the accuracy of the correlation coefficient. Also the number of candidates is small for statistical purposes, so that the coefficients obtained have rather high probable errors. For these reasons the reliabilities only of certain selected tests were calculated; these were the tests showing the least and greatest values of h, $\alpha$, and c. The results were as follows.

Table 36. Relation of reliability to h, $\alpha$, and c.

| Test | Special feature | h | $\alpha$ | c | Reliability |
|---|---|---|---|---|---|
| 27 | Smallest h | .06 | .31 | .9 | .82 |
| 41 | Greatest h | .41 | .42 | .8 | .77 |
| 39 | Greatest $\alpha$ | .19 | .50 | .9 | .70 |
| 24 | Smallest $\alpha$ | .20 | -.22 | .8 | .49 |
| 20 | | .14 | .04 | .9 | .71 |
| 27 | | .06 | .31 | .9 | .82 |
| 34 | Greatest c | .11 | .33 | .9 | .74 |
| 39 | | .19 | .50 | .9 | .70 |
| 11 | | .34 | .23 | .6 | .81 |
| 28 | | .40 | .20 | .6 | .79 |
| 32 | Smallest c | .33 | -.17 | .6 | .73 |
| 33 | | .26 | -.10 | .6 | .73 |

It is evident from the above table that there is no clear relationship between reliability and hig as measured by h, $\alpha$, or c.

(b) Data of physics tests.

In the case of this and other groups of complete tests, no values of $\alpha$ are available. The difficulty of coarseness of grouping in the correlation table is even more pronounced with these physics tests, since they were 8 item tests. It was probably on this account that some of the reliabilities obtained were so low. In one test, (F), perfect scores obtained by 4 candidates were omitted. If allowed to stand these scores would increase the reliability coefficients in an artificial way. It must also be noted that the coefficient of steepness given is calculated from a test of 8 items, and may be used only to compare the steepnesses of 8 item tests, as is done in the table.

Table 37. Relation of reliability to h and c, tests D, E, F.

| Test | $n_o$ | h | c | Reliability |
|---|---|---|---|---|
| D | 34 | .18 | .88 | .19 |
| E | 34 | .07 | .67 | .49 |
| F | 30 | .22 | .81 | .56 |

Once again the results are not very enlightening.

(c) Data of thesis tests.

Tests A, B, and C provide much more suitable material. Each test contained 15 items, so that the effects of coarseness of grouping were not so evident. The number of candidates was also reasonably large. On this occasion it was zero scores which had to be eliminated from the data, to avoid an artificial boosting of the correlation. The pmission of these scores does not affect the value of h, but the coefficient c must be recalculated, since the path from $n_o$ to $n_1$ has been altered, and with it all three lengths considered in the definition of that coefficient When these precautions had been taken the results were as follows.

Table 38. Relation of reliability to h and c, tests A, B, and C.

| Test | $n_o$ | h | c | Reliability |
|---|---|---|---|---|
| A | 118 | .40 | .63 | .77 |
| B | 159 | .09 | .76 | .75 |
| C | 160 | .07 | .86 | .86 |

The reliability of the steepest test (C) is significantly greater than that of either of the others, when the usual test is applied.

Unfortunately no reliability coefficients seem to be available for the M.H.T. group of tests. Perhaps the loss is more apparent than real, for these tests are nearly all of the steep type for the number of items they contain. This would greatly diminish their usefulness as a group in which to study the relation of hig to reliability. The only test to which this does not apply, test 12p, is also rather spoiled for comparison with the others, for it is of the flat type, which tends to increase the incidence of hig, but it has very few items, which tends to decrease the incidence of hig. Whether the one effect compensates the other we cannot say.

On the experimental evidence considered above it is impossible to decide whether the minimising of hig in a test thereby increases its reliability. The solution of this problem, as of others raised in this thesis, must await the advent of more extensive data.

PART TWO.

Notes on the Moray House Tests of Intelligence referred to in Part One, with tables of data.

These notes are intended to indicate features of interest in the Moray House series of Tests from the point of view of the preceding chapters. Through the kindness of Professor G. H. Thomson, there is included in this part a copy of each of the tests. Following each are tables of the frequencies of correct answers and the score-scatters obtained when the test was applied to specified populations.

There are points of interest common to all the tests. One is that the items have been arranged roughly in increasing order of difficulty. This fact, and the direction printed on most "Begin at the beginning, and go straight through" tend to reduce the amount of hig in all the tests. That amount, as measured by h, is low in all. A second point is that in most of the tests the answer-pattern is a straight line sloping from a very easy item to a very difficult item. With such an answer-pattern, and unig type of answering, the score-scatter would show the same number of candidates for every score in the range. Now the score-scatters produced are almost normal, or Gaussian, distributions. It follows that this normality is not caused by, but rather occurs in spite of, the nature of the answer-pattern coupled with a low degree of hig.

The tests included below are M.H.T. 8, 9, 11, 12v and 12p.

M. H. T. 8.

# DO NOT OPEN THIS BOOK UNTIL YOU ARE TOLD.

Examination Number only.

# LANCASHIRE EDUCATION COMMITTEE.

## EXAMINATION FOR JUNIOR SCHOLARSHIPS, 14th FEBRUARY, 1931.

## INTELLIGENCE TEST.

Age last Birthday.....

Date of Birthday.....

## INSTRUCTIONS.

When you are told to begin, answer the questions as quickly and as carefully as you can.

Begin at the beginning and go straight through.

If you cannot do any question in any test, leave it out and go on to the next.

When you finish one page, go on to the next.

You will have 45 minutes, and you will be told the time every quarter of an hour. No one is expected to do everything. Just do as much as you can.

ASK NO QUESTIONS AT ALL.

M.H.T. 8. 14/2/31.

·9S+88.

## TEST 1a.—FOLLOWING DIRECTIONS.

**Read each question carefully, and then write the answer to it in the bracket.**

**The alphabet is printed here to help you :—**

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

**BEGIN HERE :—**

(1) Do you understand you must not look at the paper of any other pupil during this examination ? If so, write P ... ... ... ... ... ... ... ( )

(2) Write the letter which comes before R in the alphabet ... ... ... ... ( )

(3) Write the odd numbers between 1 and 9, and underline the middle one ( )

(4) If 22 farthings are the same as $6\frac{1}{2}$ pence, write O ; if not, write C ... ... ... ( )

(5) Write the letter which occurs most often in the word IRRADIATION ... ... ( )

(6) If M and K are separated in the alphabet by one letter, write it down ; if not, write A ... ... ... ... ... ... ... ... ... ... ... ... ( )

(7) If M comes after R in the alphabet, write P ; if not, write X ... ... ... ( )

(8) If $\frac{1}{8}$ is more than $\frac{1}{7}$, write N ; if not, write Y ... ... ... ... ... ... ( )

## TEST 2a.—ANALOGIES.

Look at the first example :—

(1) finger : hand—toe : ? ... ... (foot, knee, arm, shoe, nail)

This means that finger is to hand as toe is to what ? The answer is one of the five words in the bracket. the right answer is " foot " and it is underlined, since a finger is part of a hand just as a toe is part of a foot.

Now look at Example (2) :—

(2) man : clothes— ? : fur ... ... (coat, animal, bird, skin, cloth)

This means that man is to clothes as what is to fur ? " Animal " is the right answer, because an animal wears fur just as a man wears clothes ; so " animal " is underlined.

Now look at Example (3) :—

(3) king : queen—lord : ? ... ... (princess, sister, duke, lady, prince)

In each line you are to look at the five words in the bracket, and decide which of them should go where the question mark is, and underline it. Do them just as the examples were done. Remember, you have nothing to do but to UNDERLINE ONE word in each bracket.

**BEGIN HERE :—**

(1) brother : sister—nephew : ? ... (cousin, niece, boy, girl)
(2) up : down—west : ? ... ... (north, opposite, east, over, south)
(3) ship : steamer— ? : tiger ... (animal, eagle, lion, camel, runner)
(4) king : country— ? : school ... (teacher, caretaker, scholar, prince, headmaster)
(5) adjective : noun— ? : verb ... (proverb, adverb, subject, object, preposition)
(6) lean : fat—small : ? ... ... (full, many, empty, large, much)
(7) field : gate—house : ? ... ... (window, room, chimney, door, wall)

**Go on to NEXT PAGE without waiting to be told.**

## TEST 3a.—NUMBER SERIES.

Look at the first line of numbers :—

Example (1) 1 2 3 4 5 ... ... ... ... ... ... ... ( 6 )

The one that comes next is 6, because the numbers go up one at a time. In each line there is a rule for finding the next number. In this one the rule is that the numbers go up by 1 each time. The other lines have different rules.

Example (2) 12 10 8 6 4 ... ... ... ... ... ... ... ( 2 )

Here the rule is that the numbers come down by 2 at each time.

Example (3) 1 2 4 8 16 ... ... ... ... ... ... ... ( 32 )

Here the rule is that each number is twice as big as the one before it, so the answer in the bracket is 32.

Now try the lines below. In each line find the rule, and then write the number that should come next in the bracket.

**BEGIN HERE :—**

2 5 8 11 14 17 ... ... ... ... ... ... ... ... ( )

31 28 25 22 19 ... ... ... ... ... ... ... ... ... ( )

$\frac{1}{11}$ $\frac{1}{10}$ $\frac{1}{9}$ $\frac{1}{8}$ $\frac{1}{7}$ ... ... ... ... ... ... ... ... ... ( )

28 21 14 7 ... ... ... ... ... ... ... ... ... ... ( )

4 8 12 16 12 8 ... ... ... ... ... ... ... ... ( )

3 3 5 5 7 ... ... ... ... ... ... ... ... ... ( )

2 6 18 54 ... ... ... ... ... ... ... ... ... ... ( )

## TEST 4a.—REASONING.

DIRECTIONS.—Three answers to each question are given in the bracket after it. You are to underline what you think is the RIGHT answer. You have nothing to write. Only UNDERLINE.

(1) Tom has more money than Dick, and Dick has more money than Harry. Who has the most money of the three ? (Tom, Dick, Harry)

(2) Ada is smaller than Bertha, but not so small as Clara. Who is the smallest of the three ? ... ... ... (Ada, Bertha, Clara)

(3) Wool is dearer than cotton, and silk is dearer than wool. Which is the cheapest ?... ... ... ... ... (wool, silk, cotton)

(4) Mr. Smith's house is larger than Mr. Jack's, and Mr. Watt's is larger than Mr. Smith's. Who has the largest house ? ... ... ... ... ... ... ... (Mr. Smith, Mr. Jack, Mr. Watt)

(5) Three people A, B, C, set out from London. A goes half as far as C, but twice as far as B. Who went farthest ? (A, B C)

(6) Mr. Ross's house is near the grocer's shop, but Mr. Page's house is nearer still ; while Mr. Robb's house lies between the other two. Who is nearest the grocer's shop ? (Mr. Ross, Mr. Page, Mr Robb)

## TEST 5a.—CLASSIFICATION.

Look at the first example :—

Example (1) : bullet cannon gun pencil sword

Here we have a line of five words. One of them, " pencil," has been underlined. The other four words of names of things used for fighting. But a pencil is not used for fighting, so we underline it.

Now look at the second example :—

Example (2) : grass bread meat milk potatoes

Again, we have a line of five words. " bread," " meat," " milk," and " potatoes " are the names of things we eat. But we do not eat " grass," so we underline it.

Look at the third example :—

Example (3) : mill fill pill spill say

Again, we have a line of five words. The first four—" mill," " fill," " pill," " spill," sound like each other ; but " say " does not sound like them at all. It sounds quite different, so we underline it.

Now try the following. In each line underline JUST ONE WORD that does not belong there.

| | | | | | |
|---|---|---|---|---|---|
| (1) | red | square | blue | green | yellow |
| (2) | ball | cube | circle | coin | ring |
| (3) | wood | lead | iron | copper | gold |
| (4) | Mary | Tom | Sam | Dick | John |
| (5) | bad | grand | fine | splendid | good |
| (6) | arm | skin | hair | glove | foot |
| (7) | big | huge | small | large | great |
| (8) | begin | originate | commence | start | finish |

## TEST 1b.—FOLLOWING DIRECTIONS.

**Read each question carefully, and then write the answer to it in the bracket.**

**The alphabet is printed here to help you :—**

**A B C D E F G H I J K L M N O P Q R S T U V W X Y Z**

(1) If X comes after V in the alphabet, and if L comes after P, write G ; but if only one of these is true, write M ... ... ... ... ... ... ... ... ( )

(2) The first letters of the four directions (East, etc.) when put in a certain order form a word ; write it in the bracket ... ... ... ... ... ( )

(3) HTOMMAM is a word seen in a mirror. Write it as it usually appears ( )

(4) Write P, unless the second letter of this sentence is R ; if it is, write T ... ... ( )

(5) If the letters in the alphabet were written starting from the other end, what would the 13th letter be ? ... ... ... ... ... ... ... ... ... ( )

(6) Suppose all the even letters in the alphabet came first, then the odd ones, what would the fifth letter of the alphabet then be ? ... ... ... ... ( )

(7) Write the letter which follows the letter which comes after E ... ... ... ( )

(8) If the alphabet began at K, the preceding letters being put at the end, what would the 8th letter be ? ... ... ... ... ... ... ... ... ... ( )

(9) Write the letter which is next but one after the letter between K and M ... ... ( )

**Go on to NEXT PAGE without waiting to be told.**

## TEST 2b.—ANALOGIES.

This is like Test 2a on page 2. You may look back at the directions if you wish.

You underline ONE word in each bracket.

(1) hat : head—glove : ? ... ... ... (arm, hand, elbow, foot, face)
(2) A : Z—beginning : ? ... ... ... (after, start, complete, end, distant)
(3) needle : prick—knife : ? ... ... (sharp, fork, point, blade, cut)
(4) hive : bee— ? : man ... ... (food, work, honey, meat, house)
(5) a : d — first : ? ... ... (second, fourth, later, last, third)
(6) pleasure : rejoice—doubt : ? ... ... (sorrow, lament, believe, act, hesitate)
(7) cellar : coal— ? : milk ... ... (man, tea, coffee, tub, jug)
(8) anger : pleasure—rage : ? ... ... (hesitation, delight, sorrow, lament, expect)

## TEST 3b.—NUMBER SERIES.

This is like Test 3a on page 3. You may look back at the directions if you wish.

Write the number that comes next in the bracket.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\frac{1}{2}$ | $\frac{1}{5}$ | $\frac{1}{8}$ | $\frac{1}{11}$ | $\frac{1}{14}$ | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 3 | $4\frac{1}{2}$ | 6 | $7\frac{1}{2}$ | 9 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| $5\frac{1}{4}$ | $8\frac{1}{4}$ | $11\frac{1}{4}$ | $14\frac{1}{4}$ | $17\frac{1}{4}$ | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 4 | 6 | 9 | 13 | 18 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 2 | 6 | 12 | 20 | 30 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 1 | 4 | 16 | 64 | 256 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 20 | 19 | 17 | 14 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |

## TEST 4b.—REASONING.

**Remember that you underline the RIGHT answer.**

(1) John is better than Tom at composition. But Tom is better than John at drawing. If composition is more important than drawing, who is the better of the two ? (John, Tom, No one can tell)

(2) Tom, John and Peter are sitting in a row. John is on the left of Peter, and Tom is to the left of John. Who is in the middle ? ... ... ... ... ... (John, Tom, Peter)

(3) If Dick is put on the right of Harry, and Andrew is put on the right of Dick, who is now in the middle ? ... (Dick, Andrew, Harry)

(4) All kinds of wood float on water. A piece of material is thrown on water, and it floats. Is it wood ? ... (Yes, No, No one can tell)

(5) "Umo" is twice as dear as "Ritka"; "Ritka" is twice as good for food as "Umo" is. I have 1/- to spend. Is it better to buy "Umo" or "Ritka" ? ... ("Umo", "Ritka", No one can tell)

## TEST 5b.—CLASSIFICATION.

Remember that in each line there is one word that does not belong there ; it is different in some way from the others.

When you have found it, underline it. Underline JUST ONE WORD in each line.

| | | | | | |
|---|---|---|---|---|---|
| (1) | jump | leap | halt | walk | run |
| (2) | receive | give | find | beg | borrow |
| (3) | cloth | butter | bread | cake | beef |
| (4) | hurt | pleasure | enjoyment | gladness | ease |
| (5) | mend | create | construct | destroy | produce |
| (6) | Cæsar | Napoleon | Milton | Haig | Wellington |
| (7) | pencil | ink | pen | chalk | crayon |
| (8) | gramophone | violin | sonata | piano | organ |
| (9) | assemble | gather | amass | collect | disperse |

## TEST 1c.—FOLLOWING DIRECTIONS.

**Remember you write the answer to the question in the bracket.**

**A B C D E F G H I J K L M N O P Q R S T U V W X Y Z**

(1) If the word FACETIOUS contains the vowels in their proper order, write L ; if not, write P ... ... ... ... ... ... ... ... ... ... ( )

(2) Write the two letters in the word SMALLER that have as many letters between them in the alphabet as there are letters in the word PORT ... ... ( )

(3) Write X if SUBSEQUENT contains the 17th letter of the alphabet, unless F comes in the alphabet after the second letter of the word CHART, in which case write C ... ... ... ... ... ... ... ... ... ... ( )

(4) If the letters in the word BEGINS appear in the order in which they are found in the alphabet, write the letter which is midway in the alphabet between the second and sixth letters in the word ; if not, write O ... ... ... ( )

(5) Harry said over the letters of the alphabet till he came to M ; then went backwards for six letters ; then forward again for two letters. What was the letter to the left of that at which he stopped ? ... ... ... ... ... ... ( )

(6) If the letters in the word YES appear in the same order as they do in the alphabet ; and if the same is true for the letters of the word NO, write X ; but if this is true of only one of the words, write T ... ... ... ... ... ... ( )

(7) If each pair of letters in the alphabet were interchanged, so that it now read BADC . . . . , what would the 13th letter in the alphabet be then ? ... ( )

(8) Write the letter in the alphabet midway between the two letters which occur most often in the word VICISSITUDES ... ... ... ... ... ... ( )

## TEST 2c.—ANALOGIES.

**Remember you underline the right word in each bracket.**

(1) air : breathing—water : ? ... (swimming, floating, washing, drinking)
(2) calculate : reason—anger : ? ... (thought, feeling, imagination, memory)
(3) hand : foot—finger : ? ... ... (leg, ankle, toe, palm, elbow)
(4) petrol : car— ? : train ... ... (wheels, smoke, engine, steam)
(5) popular : applause—criminal : ?... (punishment, misery, reward, judgment

**Go on to NEXT PAGE without waiting to be told.**

## TEST 3c.—NUMBER SERIES.

**Remember that you write the number that should come next in the bracket.**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 6 | 12 | 20 | 30 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 5 | 10 | 20 | 40 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 1 | 4 | 2 | 5 | 3 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 2 | 7 | 14 | 23 | 34 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 2 | 3 | 2 | 4 | 2 | 5 | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 7 | 1 | 6 | 2 | 5 | 3 | 4 | ... | ... | ... | ... | ... | ... | ( ) |
| 600 | 300 | 100 | 25 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 625 | 125 | 25 | 5 | 1 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |

## TEST 4c.—REASONING.

**Remember that you underline the right answer.**

(1) As I stand with my back to the rising sun, my house is on my left hand. In what direction must I walk to get home ? (North, South, West)

(2) A poet writes three poems. The first has three verses of four lines each ; the second has two verses of six lines each ; the third has five verses of two lines each. Which is the shortest poem ; ... ... ... ... ... (first, second, third)

(3) Three cars travel along a road. A passes B but cannot pass C. The one now at the back increases its speed, and passes the other two. Which is now farthest behind ? ... ... ... ... ... ... ( A, B, C )

(4) There is a town, all of whose streets run either North and South, or East and West. Walking along Brown Street, I am going East. I turn to the left along Hillside Street ; then to the right along London Street, then to the left along Oxford Street. In what direction does London Street run ? ... ... ... ... ... ... (North and South
East and West
No one can tell)

(5) Mrs. Smith, Mrs. Brown and Mrs. Jones buy the same kind of cloth at the same shop. Mrs. Smith buys much more than Mrs. Brown, who, however, spends a little less than Mrs. Jones. Their daughters, Miss Smith, Miss Brown and Miss Jones buy a hat each, and the account paid by each mother for herself and her daughter is the same. Which of the daughters chose the most expensive hat ? ... ... ... ... (Miss Smith, Miss Brown,
Miss Jones)

## TEST 5c.—CLASSIFICATION.

Remember that you underline the word in each line that does not belong there.

Cross out JUST ONE WORD in each line.

| | | | | | |
|---|---|---|---|---|---|
| (1) | crystal | wall | spectacles | bottle | window |
| (2) | telephone | tramcar | steamer | train | cab |
| (3) | Raphael | Collie | Spaniel | Terrier | Pomeranian |
| (4) | wheat | oats | turnips | rye | barley |
| (5) | bullet | knife | spoon | book | key |
| (6) | support | hinder | assist | help | encourage |
| (7) | red | violet | yellow | pansy | blue |
| (8) | explain | show | tell | describe | narrate |
| (9) | fairest | rarest | carest | farthest | greatest |

The frequencies of correct answers given below were obtained from the papers of 528 candidates, and the score-scatter from those of 6423 candidates. For these figures the author is indebted to the Director of Education of the county where this test was tried out. A first estimate of whether the 528 candidates are representative of the total 6423 can be made by comparing their average scores, which were 71 for the sample and 74 for the whole group.

The frequencies of correct answers by the 528 candidates were given as percentages, and are tabulated as such. The answer-pattern and score-scatter are graphed after the tables.

In all these tests it may be taken for granted that the average age of the testees is eleven years.

Table 39. Percentages of correct answers in M.H.T. 8.

| | Subtests | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | 1a | 2a | 3a | 4a | 5a | 1b | 2b | 3b | 4b | 5b | 1c | 2c | 3c | 4c | 5c |
| 1 | 98 | 92 | 92 | 93 | 97 | 90 | 94 | 89 | 70 | 91 | 62 | 58 | 62 | 32 | 53 |
| 2 | 96 | 90 | 89 | 91 | 60 | 51 | 81 | 86 | 82 | 18 | 26 | 44 | 60 | 74 | 66 |
| 3 | 44 | 53 | 94 | 91 | 89 | 60 | 76 | 84 | 80 | 94 | 61 | 61 | 60 | 53 | 54 |
| 4 | 95 | 51 | 68 | 88 | 83 | 75 | 81 | 75 | 44 | 88 | 9 | 52 | 42 | 20 | 68 |
| 5 | 88 | 74 | 88 | 74 | 81 | 89 | 55 | 63 | 55 | 49 | 38 | 45 | 57 | 27 | 41 |
| 6 | 85 | 88 | 91 | 87 | 87 | 41 | 42 | 60 | - | 48 | 83 | - | 33 | - | 52 |
| 7 | 62 | 64 | 71 | - | 66 | 73 | 86 | 75 | - | 49 | 42 | - | 38 | - | 57 |
| 8 | 79 | - | - | - | 63 | 70 | 72 | - | - | 58 | 42 | - | 19 | - | 10 |
| 9 | - | - | - | - | - | 64 | - | - | - | 71 | - | - | - | - | 12 |

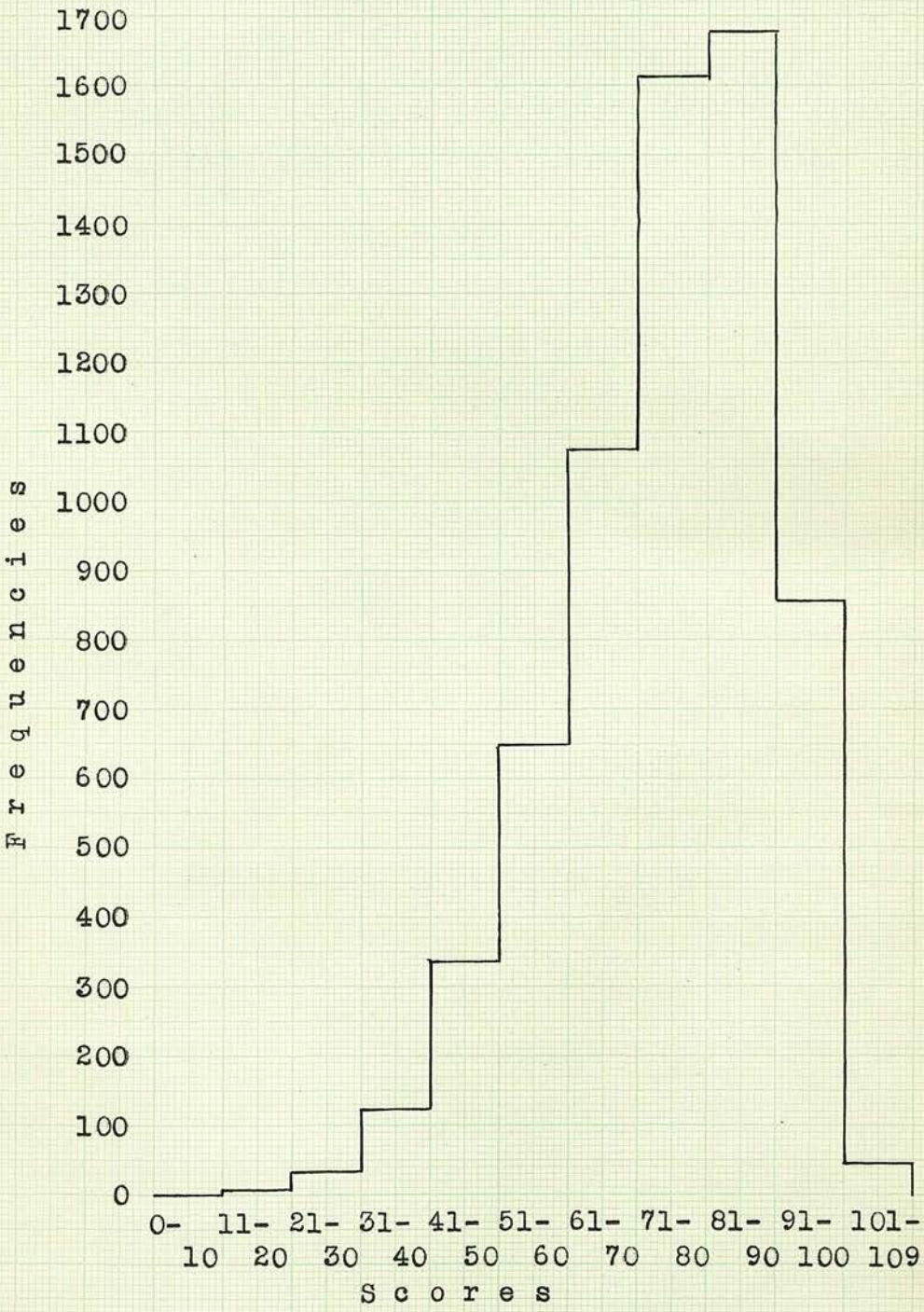
Table 40. Score-scatter of M.H.T. 8.

| Score | Frequency |
|---|---|
| 0 - 10 | 0 |
| 11 - 20 | 9 |
| 21 - 30 | 33 |
| 31 - 40 | 124 |
| 41 - 50 | 337 |
| 51 - 60 | 650 |
| 61 - 70 | 1076 |
| 71 - 80 | 1613 |
| 81 - 90 | 1679 |
| 91 -100 | 857 |
| 101 -109 | 45 |
| | 6423 |

This is a noteworthy case of the score-scatter being skewed negatively by a negatively skewed answer-pattern-differential. The skewness of the answer-pattern-differential is -0.38, and that of the score-scatter is -0.66 . The coefficient of hig is 0.11 .

Figure 26. Answer-pattern of M.H.T. 8.
( from 528 candidates )

Figure 27. Score-scatter of M.H.T. 8.
6423 candidates.



M. H. T. 9.

NOT TO BE FILLED IN BY THE SCHOLAR.

| Age (years and months) on 1/8/31. | Raw Score. | I.Q. |
|---|---|---|
| | | |

# County Borough of Halifax—Education Committee.

# INTELLIGENCE TEST, 1931.

Write your Surname here.....

Christian name here .....

School .....

What standard or class or form are you in ?.....

## INSTRUCTIONS.

When you are told to begin, answer the questions as quickly and as carefully as you can.

Begin at the beginning and go straight through.

If you cannot do any question in any test, leave it out and go on to the next.

When you finish one page, go on to the next. Be sure you do not turn over two pages at once.

You will have 45 minutes. You will be told the time every quarter of an hour.

*ASK NO QUESTIONS AT ALL.*

## TEST 1a.—FOLLOWING DIRECTIONS.

**Read each question carefully, then write the answer to it in the bracket.**

The alphabet is printed here to help you :—

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

**BEGIN HERE :—**

Do you understand that you must do your best and not ask questions ? If so, write M ( )

Have you noticed that there is an alphabet printed near the top of this page to help you ?
If so, write the 5th letter in the alphabet ... ... ... ... ... ... ( )

Write the letter before O in the alphabet ... ... ... ... ... ... ... ( )

Write the letter which occurs most often in the word UNDERSTANDING ... ... ( )

If the alphabet were written backwards, starting with Z, what would the 7th letter be ? ( )

Write the numbers between 4 and 9 and underline the smallest one ... ( )

If the letters K and L changed places in the alphabet, what would the 12th letter be ? ( )

If there are 14 farthings in $2\frac{1}{2}$ pence, write Y ; if not, write N ... ... ... ... ( )

## TEST 1b.—NUMBER SERIES.

Look at the first line of numbers :—

Example (1) 1 2 3 4 5 ... ... ... ... ... ... ... ( 6 )

The one that comes next is 6, because the numbers go up one at a time. In each line of numbers below there is a rule for finding the next number. In this one the rule is that the numbers go up by 1 each time. The other lines have different rules.

Example (2) 12 10 8 6 4 ... ... ... ... ... ... ... ( 2 )

Here the rule is that the numbers come down by 2 each time.

Example (3) 1 2 4 8 16 ... ... ... ... ... ... ... ( 32 )

Here the rule is that each number is twice as big as the number before it, so the answer in the bracket is 32.

Now try the lines below. In each line find the rule, and then write in the bracket the number that should come next.

2 4 6 8 10 12 ... ... ... ... ... ... ... ... ( )

3 7 11 15 19 ... ... ... ... ... ... ... ... ... ( )

2 2 3 3 4 ... ... ... ... ... ... ... ... ... ( )

1 $\frac{1}{3}$ $\frac{1}{5}$ $\frac{1}{7}$ $\frac{1}{9}$ ... ... ... ... ... ... ... ... ... ( )

16 13 10 7 4 ... ... ... ... ... ... ... ... ... ( )

2 6 18 54 162 ... ... ... ... ... ... ... ... ( )

## TEST 1c.—ANALOGIES.

Look at the first example :—

(1) finger : hand—toe : ? ... ... (foot, knee, arm, shoe, nail)

This means that finger is to hand as toe is to what ? The answer is one of the five words in the bracket. The right answer is "foot," so it is underlined ; it is the right answer, since a finger is a part of a hand, just as a toe is part of a foot.

Now look at Example (2) :—

(2) man : clothes— ? : fur ... ... (coat, animal, bird, skin, cloth)

This means that man is to clothes as what is to fur ? Now a man wears clothes just as an animal wears fur, so "animal" is the correct answer, and is therefore underlined.

Now look at Example (3) :—

(3) king : queen—lord : ? ... ... (princess, sister, duke, lady, prince)

In each line below you have to look at the five words in the bracket, decide which should go where the question mark is, and underline it. All you have to do is UNDERLINE ONE word in each bracket.

**BEGIN HERE :—**

horse : animal—swallow : ? ... ... (summer, fly, nest, bird, swift)
gate : field—door : ? ... ... ... (window, room, grass, stile, hinge)
feathers : hen—wool : ? ... ... (duck, jersey, sheep, blanket, coat)
arm : wrist—leg : ? ... ... ... (knee, elbow, ankle, bones, foot)
woman : girl— ? : boy ... ... (father, man, lad, youth, nurse)
long : short— ? : poor ... ... (thin, happy, small, content, rich)
before : after— ? : now ... ... (soon, then, but, why, if)
wing : bird— ? : fish ... ... ... (tail, mouth, fin, sea, feathers)

## TEST 1d.—REASONING.

DIRECTIONS.—Three answers to each question are given in the bracket after it. You are to underline what you think is the RIGHT answer. You have nothing to write. Only UNDERLINE ONE answer in each bracket.

**BEGIN HERE :—**

Segrave's car is faster than Campbell's car, and Campbell's car is faster than Harvey's. Who has the fastest car ? (Segrave, Campbell, Harvey)

Iron is stronger than wood, but not so strong as steel. Which is the strongest ? ... ... ... ... ... ... (iron, wood, steel)

John is taller than Tom, and Harry is taller than John. Who is the tallest ? ... ... ... ... ... ... (John, Tom, Harry)

Mary is older than Jane, and Ella is younger than Mary. Which of the three girls is the oldest ? ... ... (Mary, Jane, Ella)

I have three cricket bats. The first is heavier than the second, and the second is heavier than the third. Which is the lightest ? ... ... ... ... ... ... ... (first, second, third)

Three sticks of different lengths are coloured, one red, one blue, one green. The blue one is longer than the red one, and the green one is shortest of all. Of what colour is the longest stick ? ... ... ... ... ... ... (red, blue, green)

## TEST 1e.—LOGICAL SELECTION.

Look at this :—

dog ... (collar, hair, muzzle, chain, legs)

A dog ALWAYS has hair and legs, so these are underlined in the bracket. It does not always have a collar, or a muzzle, or a chain, so these are not underlined. In the lines below underline the TWO words which tell what the thing outside the bracket always has. Remember to UNDERLINE TWO words only in each bracket.

**BEGIN HERE :—**

boy ... (head, jacket, skin, hat, boots)

horse ... (stable, mouth, saddle, hoof, shoe)

motor car (petrol, smoke, engine, noise, wheels)

river ... (fish, banks, bridge, boat, water)

room ... (pictures, window, door, wall, table)

race ... (start, runners, spectators, competitors, prize)

## TEST 2a.—FOLLOWING DIRECTIONS.

**Read each question carefully, and then write the answer to it in the bracket.**

**The alphabet is printed here to help you :—**

**A B C D E F G H I J K L M N O P Q R S T U V W X Y Z**

If S is found before U in the alphabet, and M before K, write Y, but if only one of these is true, write C ... ... ... ... ... ... ... ... ... ... ( )

Four months of the year have names ending in the same three letters. Write the middle letter of these three ... ... ... ... ... ... ... ... ... ( )

Another four months have names ending in the same letter (not the same letter as in the previous question.) Write this letter ? ... ... ... ... ... ... ( )

If the alphabet had only 24 letters, F and G being dropped out, what would the 15th letter be ? ... ... ... ... ... ... ... ... ... ... ( )

OTTOM is a word seen in a mirror. Write it as it usually appears ... ( )

Write N unless the last letter of this sentence is R, in which case write that letter ... ( )

If the alphabet began with J, would the 10th letter be T ? If so, write A ; If not, write B ... ... ... ... ... ... ... ... ... ... ... ( )

If there are as many letters between H and M in the alphabet as there are between U and P, write Z ; if not, write X ... ... ... ... ... ... ... ( )

## TEST 2b.—NUMBER SERIES.

This is like Test 1b on page 2. You may look back at the directions if you wish. Write in the bracket the number that comes next.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 9 | 27 | 81 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 2 | 4½ | 7 | 9½ | 12 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 1 | 2 | 4 | 7 | 11 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 25 | 20 | 16 | 13 | 11 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 5½ | 9½ | 13½ | 17½ | 21½ | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 0 | 3 | 8 | 15 | 24 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 3 | 6 | 12 | 15 | 30 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |

**Go on to NEXT PAGE without waiting to be told.**

## TEST 2c.—ANALOGIES.

This is like Test 1c on page 3. You may look back at the directions if you wish.

You UNDERLINE ONE word in each bracket.

coal : fire—food : ? ... ... ... (plate, stomach, knife, shop, pot)

square : circle—cube : ? ... ... (triangle, oblong, sphere, block, line)

wax : wane—stretch : ? ... ... (burst, swell, shape, shrink, round)

rail : wheel— ? : foot ... ... (leg, ground, spoke, hand, shoe)

bird : cage— ? : kennel ... ... (straw, chain, roof, dog, cat)

vegetable : carrot— ? : banana ... (yellow, skin, fruit, apple, red)

decision : hesitation—certainty : ? ... (thought, perplexity, assurance, judgment, relief)

contempt : admiration— ? : admire... (despise, condemn, ridicule, loathe, ignore)

## TEST 2d.—REASONING.

**Remember that you underline the right answer.**

A wire fence is supported by posts, spaced at a distance of one yard from each other. If there are twenty such posts on this fence, what is the distance in yards from the first post to the twentieth one? ... ... ... ... (nineteen, twenty, twenty-one)

Pit ponies are said to become blind through working underground. John has a pony which is blind. Was it formerly a pit pony? ... ... ... ... ... (Yes, No, I cannot tell)

In Willie's home there are his father and mother, his two sisters, and one brother. How many males are there in the household? ... ... ... ... ... ... (one, two, three)

How many daughters has Willie's father? ... ... ... (one, two, three)

Mr. Wilson is not healthy, and cannot travel for more than three hours at a stretch. He also feels sick if he is in a train for more than two hours, or in an aeroplane for more than one. If a motor car travels at 30 miles an hour, a train at 50 miles an hour, and an aeroplane at 80 miles an hour, which should Mr. Wilson use for a non-stop journey of 95 miles? ... ... ... ... (motor car, train, aeroplane)

In a foreign seaport, an Englishman who could speak only English wished to speak to a Chinaman, who knew only Chinese. The Englishman obtained a Frenchman who spoke both French and English, and a Russian who could speak both Russian and French. Meanwhile, the Chinaman met a fellow Chinaman who could speak both Russian and Chinese. Would the Englishman now be able to converse with the Chinaman?... ... ... (Yes, No, I cannot tell.

## TEST 2e.—LOGICAL SELECTION.

Underline the TWO words in the bracket which tell what the thing outside the bracket is certain, or most likely, to have or to be connected with. This is like Test 1e on Page 4.

ship ... (sails, engine, hull, funnel, rudder)

enthusiasm (energy, patience, thought, zeal, nobility)

respect ... (malice, hatred, obedience, love, envy)

field ... (grass, earth, hedge, area, trees)

journey ... (return, departure, route, result, object)

## TEST 3a.—FOLLOWING DIRECTIONS.

**Remember to write the answer to the question in the bracket.**

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

If the alphabet were written backwards, write the letter which would be 8th in the new order ... ... ... ... ... ... ... ... ... ... ... ( )

If, after the alphabet written backwards, the alphabet was written again in the correct way, which letter would have most letters between its two appearances? ... ( )

Write the letter which would have six letters between its two appearances ... ... ( )

If there are more I's in DIMINISHING than in TRINITARIAN write P, unless there are more N's in the second than in the first, in which case write R ... ... ( )

If the letters of the word TROPICAL occur in the order opposite to their order in the alphabet, write O, but if not, put them in that order, and write here the letter which has to be shifted farthest ... ... ... ... ... ... ... ( )

If the 2nd, 4th, 6th and all the other even letters of the alphabet were lost, write what would remain of the word RINTINTIN ... ... ... ... ( )

## TEST 3b.—NUMBER SERIES.

**Remember to write in the bracket the number that should come next.**

624 312 156 78 ... ... ... ... ... ... ... ... ... ... ( )

19 18 20 19 21 ... ... ... ... ... ... ... ... ... ( )

4 8 16 32 ... ... ... ... ... ... ... ... ... ... ( )

216 36 6 1 ... ... ... ... ... ... ... ... ... ... ( )

1 3 6 10 15 ... ... ... ... ... ... ... ... ... ( )

63 42 25 12 ... ... ... ... ... ... ... ... ... ... ( )

1 $\frac{3}{2}$ $\frac{9}{4}$ $\frac{27}{8}$ $\frac{81}{16}$ ... ... ... ... ... ... ... ... ( )

1 4 8 11 22 ... ... ... ... ... ... ... ... ... ( )

## TEST 3c.—ANALOGIES.

**Remember to underline the right word in each bracket.**

sacred : secular— ? : hall ... ... (mansion, church, window, cemetery, door)

uncle : nephew—aunt : ? ... ... (uncle, daughter, cousin, niece, son)

receipt : bill— ? : debt ... ... (account, credit, transaction, payment, money)

advance : retire—ascend : ? ... ... (climb, hill, withdraw, descend, retreat)

evolution : revolution—fire : ?... ... (smoke, burning, explosion, heat, matches)

migration : birds— ? : words ... (syllables, alteration, dictionary, translation, sentence)

## TEST 3d.—REASONING.

**Remember that you underline the right answer.**

Mr. Jones' watch always stopped when the hands passed each other, and then he had to restart it. If it was going at 6 a.m., how many times would he have to restart it by 6 p.m. ... ... ... ... ... ... ... (eleven, twelve, thirteen)

Mr. Brown is middle aged and can walk three miles in an hour. His son is young and can walk four miles in an hour, while his father, the son's grandfather, is so old that he can walk only two miles in an hour. The three go out together for an hour's walk. How far will they go? (two miles, three miles, four miles)

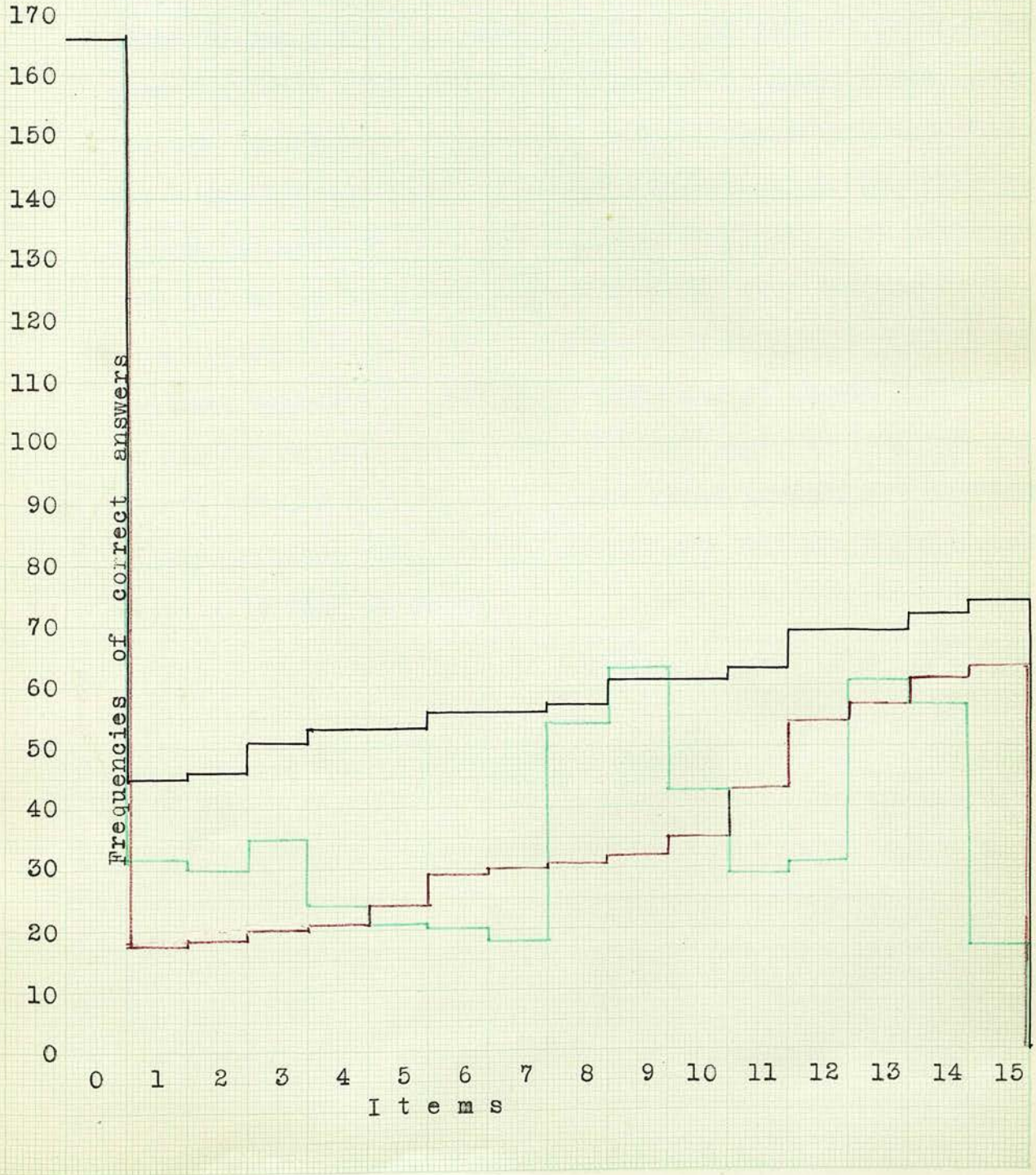
A Wolf Cub gains a star for each year he has been a Cub. Johnny who has been a Cub for nearly four years found a jersey of his which had one star on it. What is the least number of years he can have had this jersey? ... (one, two, three)

Mr. Brown came home from holiday and left the key of the home letter box at the seaside with his wife. He asked her to post it to him, and she did. Was that sensible, or silly ... ... ... ... ... ... ... (sensible, silly, I cannot tell)

John, Tom and Dick all like reading. John likes detective stories and school stories; Tom likes school stories and exploration stories; and Dick likes detective stories and exploration stories. The cleverest of these boys does not read many detective stories. Which is he? (John, Tom, Dick)

**THE END. Look over your work again.**

Fig. 12. Frequencies of correct answers in Thesis Test A



Conclusion.

These results show that the answer-pattern of a test does possess a large degree of permanence, if certain rather obvious conditions are observed. If a pattern derived from the answers given by one group of candidates is to be used to predict the answer-pattern that will be produced by a second group, then the age (or ability) difference between the groups must be small. For a small difference the answer-pattern may be adjusted by the method illustrated, but for a large difference this method has been proved to be inadequate. Secondly, if a test is to be constructed from some of the items of a test with a known answer-pattern, then the test-paper so constructed should have a format as similar as possible to the original test, so that the items have a similar environment. Under these conditions, it seems highly probable that the answer-pattern obtained will correspond closely to that intended.

## Chapter 5. The Relation of Answer-pattern and Score-scatter in the General Case.

In the special case considered in chapter 2, there is an exact relationship between the answer-pattern and the score-scatter, expressed by the equations A and B of that chapter. The necessary condition was that each candidate's score be made up of answers to the easiest questions, and this state of affairs was described as unig. Tests of this type are very rare; a certain amount of hig enters into the composition of almost every test, and the exactness of the relationship between answer-pattern and score-scatter is destroyed. Although the exact relationship has gone, it may be that there persists some measure of correlation, or, to express the idea in a different way, there may stll be some degree of control of the score-scatter by the answer-pattern.

As examples, consider the following answer-patterns with their answer-pattern-differentials and score-scatters, as obtained in actual examinations. The first two are selected more or less at random from the 41 tests, of which they are tests 32 and 7, and the third from data of M.H.T. 12p already used.

Table 12. Data of test 32, test 7, and M.H.T. 12p .

Test 32

| Item or score. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Answer-pattern. | 32 | 31 | 30 | 30 | 27 | 27 | 21 | 19 | 19 | 4 | 2 |
| Answer-pattern-differential. | 1 | 1 | 0 | 3 | 0 | 6 | 2 | 0 | 15 | 2 | 2 |
| Score-scatter. | 0 | 0 | 0 | 0 | 3 | 7 | 3 | 11 | 5 | 2 | 1 |

Test 7.

| Item or score. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Answer-pattern. | 32 | 32 | 27 | 20 | 18 | 17 | 17 | 17 | 14 | 13 | 11 |
| Answer-pattern-differential. | 0 | 5 | 7 | 2 | 1 | 0 | 0 | 3 | 1 | 2 | 11 |
| Score-scatter. | 0 | 1 | 3 | 4 | 5 | 3 | 2 | 1 | 7 | 3 | 3 |

M.H.T. 12p.

| Item or score. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Answer-pattern. | 450 | 431 | 416 | 402 | 399 | 389 | 386 | 268 | 205 | 179 |
| Answer-pattern-differential. | 19 | 15 | 14 | 3 | 10 | 3 | 118 | 63 | 26 | 179 |
| Score-scatter. | 10 | 9 | 11 | 13 | 14 | 25 | 69 | 97 | 108 | 94 |

It will be readily seen that corresponding frequencies in these tables are far from being equal. Pearson's test of goodness of fit would at once declare the answer-pattern-differential and the score-scatter to be different distributions; there is not the faintest possibility that the deviations are due to errors of sampling. Yet the data seem to indicate that there is still some control of the score-scatter by the answer-pattern-differential.

For example, the two-humped answer-pattern-differential of test 7 is accompanied by a distinctly double-humped score-scatter. The positively skewed answer-pattern-differential of M.H.T. 12p is accompanied by a positively skewed score-scatter.

There are two parameters of the two distributions that are necessarily identical. The total frequency in both answer-pattern -differential and score-scatter is the same; it is $n_o$, the number of candidates. Also the first moment of both distributions is the same. In the case of the answer-pattern-differential it is $\sum_o^m x(n_x - n_{x+1})$ which equals $n_1 + n_2 + n_3 + \ldots + n_m$; that is, the number of points scored. For the score-scatter the first moment is $\sum_o^m xN_x$, which again equals the number of points scored. Expressed in another way, these latter equations mean that both distributions have the same mean. This is quite independent of the presence or absence of hig.

This line of thought suggests that with a sufficient number of tests given to the same candidates we might correlate corresponding moments. Alternatively, instead of correlating the second moments we might correlate the standard deviations; and for the third moments calculation we might substitute correlation of skewness. It is hardly necessary to proceed further than the third moment for two reasons. First, the third is the highest moment with which an examiner is ordinarily concerned, the mean score being calculable from the first moment, the standard deviation from the second and first, and the skewness from the

first three moments. Second, it is a fact well known to statisticians that spurious deviations sometimes occur in an extensive set of observations; though their effect may be small on the first, second, or even third moments, it increases rapidly with higher moments and may outweigh the sum of all the other terms, so that the calculated moment becomes quite unreliable.

The correlation of standard deviations and of coefficients of skewness was one of the main lines of attack in this investigation. The correlation of the standard deviations of answer-pattern-differential and score-scatter is reported in chapter 6, and the correlation of the measures of skewness in chapter 7.

In the case of a single test, the estimation of the degree of relationship between the answer-pattern-differential and the score-scatter is not so easy. Several methods have been tried by the author, but none has proved entirely satisfactory. The subject will be more fully discussed and the coefficients devised will be described and criticised in chapter 8.

## Chapter 6. The Correlation of the Standard Deviations of Answer-pattern-differential and Score-scatter.

### 1. Theory of the method used.

The term standard deviation was first used in the study of normal frequency distributions. If a variate x is normally distributed then two statistics, the mean and the standard deviation, summarise our knowledge of the distribution. The standard deviation, $\sigma$, measures the extent to which the variate x is scattered about the mean a.

This theory may be applied readily enough to the score-scatters found in the tests, for they approximate to normal frequency distributions, but such is not the case with the answer-pattern-differentials. These latter belong to none of Pearson's types of frequency curves. However, it seems justifiable to apply the same method of calculation to find for the answer-pattern-differential a pseudo- standard deviation, on the ground that the statistic so calculated is to be used solely for measuring the spread or scatter of the observations from the mean; no property of standard deviations which applies only to their use in normal distributions is employed.

In partial support of this argument, there may be quoted the opinion of Bowley: "Even when a group has no very close connection with the first or second approximations to the curve of error, it seems probable that c ( $=\sqrt{2}\sigma$) and j ( $=\frac{1}{2\sqrt{2}}S$ ), calculated by the methods which for the particular group seem the

least liable to chance disturbance, are the best single measures of the groupings about the average and the skewness that we can devise." ( Elements of Statistics, page 331.)

2. Experimental Results.

Very suitable data for measuring the correlation of the two standard deviations are furnished by the 41 tests. They were attempted by the same candidates throughout, and had each 10 items, avoiding any heterogeneity due to differing numbers of items.

The method of calculation is shown below; the test chosen as an example is number 4 of the 41 tests. ( See table 13 overleaf)

A provisional mean is chosen at $x = 4$. The first and second moments about that mean are calculated; if these are den denoted by $m_1$ and $m_2$ it is easy to show that $\sigma^2 = m_2 - m_1^2$. A check is provided by the fact already mentioned that the two distributions have the same first moment. In addition all the results were checked by repeating the calculations with a fresh provisional mean.

Table 13. Standard deviations of answer-pattern-differential and score-scatter.

1. Answer-pattern-differential.

| Item | A.P. | A.P.D. | x | fx | $fx^2$ |
|---|---|---|---|---|---|
| 0 | 32 | 3 | -4 | -12 | 48 |
| 1 | 29 | 2 | -3 | - 6 | 18 |
| 2 | 27 | 4 | -2 | - 8 | 16 |
| 3 | 23 | 5 | -1 | - 5 | 5 |
| 4 | 18 | 6 | 0 | - 0 | 0 |
| 5 | 12 | 3 | 1 | 3 | 3 |
| 6 | 9 | 1 | 2 | 2 | 4 |
| 7 | 8 | 1 | 3 | 3 | 9 |
| 8 | 7 | 1 | 4 | 4 | 16 |
| 9 | 6 | 0 | 5 | 0 | 0 |
| 10 | 6 | 6 | 6 | 36 | 216 |
| Total marks.145 | | $n_0$=32 | | +17 | 335 |

$m_1 = 17/32 = +0.53$ $\quad m_2 = 335/32 = 10.47$

$$\sigma_{A.P.D.} = \sqrt{m_2 - m_1^2} = \underline{3.19}$$

2. Score-scatter.

| Item | N | x | Nx | $Nx^2$ |
|---|---|---|---|---|
| 0 | 1 | -4 | -4 | 16 |
| 1 | 1 | -3 | -3 | 9 |
| 2 | 1 | -2 | -2 | 4 |
| 3 | 10 | -1 | -10 | 10 |
| 4 | 5 | 0 | 0 | 0 |
| 5 | 4 | 1 | 4 | 4 |
| 6 | 4 | 2 | 8 | 16 |
| 7 | 2 | 3 | 6 | 18 |
| 8 | 2 | 4 | 8 | 32 |
| 9 | 2 | 5 | 10 | 50 |
| 10 | 0 | 6 | 0 | 0 |
| | $n_0$..32 | | +17 | 159 |

$m_1 = 17/32 = +0.53$ $\quad m_2 = 159/32 = 4.98$

$$\sigma_N = \sqrt{m_2 - m_1^2} = \underline{2.17}$$

Check: 32×4 + 17 = 145 = total marks scored, as in answer-pattern.

Table 14. Standard deviations of answer-pattern-differential and of score-scatter of the 41 tests.

| Test | $\sigma_{A.P.D.}$ | $\sigma_{N.}$ |
|---|---|---|
| 1 | 3.11 | 2.26 |
| 2 | 3.38 | 2.26 |
| 3 | 3.64 | 2.35 |
| 4 | 3.19 | 2.17 |
| 5 | 3.37 | 2.18 |
| 6 | 3.01 | 1.96 |
| 7 | 3.79 | 2.71 |
| 8 | 2.74 | 2.08 |
| 9 | 3.08 | 1.93 |
| 10 | 3.24 | 2.22 |
| 11 | 3.84 | 2.46 |
| 12 | 3.61 | 2.11 |
| 13 | 2.89 | 2.13 |
| 14 | 3.42 | 2.29 |
| 15 | 3.89 | 2.17 |
| 16 | 2.90 | 1.90 |
| 17 | 2.54 | 1.73 |
| 18 | 3.67 | 2.32 |
| 19 | 3.61 | 2.31 |
| 20 | 3.21 | 2.11 |
| 21 | 3.20 | 1.92 |
| 22 | 3.34 | 1.94 |
| 23 | 2.86 | 1.77 |
| 24 | 2.99 | 1.75 |
| 25 | 3.33 | 2.05 |
| 26 | 3.16 | 1.97 |
| 27 | 2.92 | 2.18 |
| 28 | 3.93 | 2.37 |
| 29 | 3.35 | 2.40 |
| 30 | 3.08 | 2.02 |
| 31 | 2.70 | 1.99 |
| 32 | 2.47 | 1.52 |
| 33 | 2.70 | 1.49 |
| 34 | 2.88 | 2.09 |
| 35 | 3.11 | 2.03 |
| 36 | 3.42 | 2.10 |
| 37 | 3.51 | 2.38 |
| 38 | 3.79 | 2.51 |
| 39 | 3.56 | 2.48 |
| 40 | 3.09 | 2.09 |
| 41 | 3.43 | 2.28 |
| Means | 3.25 | 2.12 |

Table 15. Correlation of standard deviations of answer-pattern-differential and score-scatter. $r = .789$.

$\sigma_{A.P.D.}$

| $\sigma_N$ | 2.4 to 2.5 | 2.5 to 2.6 | 2.6 to 2.7 | 2.7 to 2.8 | 2.8 to 2.9 | 2.9 to 3.0 | 3.0 to 3.1 | 3.1 to 3.2 | 3.2 to 3.3 | 3.3 to 3.4 | 3.4 to 3.5 | 3.5 to 3.6 | 3.6 to 3.7 | 3.7 to 3.8 | 3.8 to 3.9 | 3.9 to 4.0 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2.7 to 2.8 | | | | | | | | | | | | | | 1 | | | 1 |
| 2.6 to 2.7 | | | | | | | | | | | | | | | | | 0 |
| 2.5 to 2.6 | | | | | | | | | | | | | | 1 | | | 1 |
| 2.4 to 2.5 | | | | | | | | | | 1 | | 1 | | | 1 | | 3 |
| 2.3 to 2.4 | | | | | | | | | | | | 1 | 3 | | | 1 | 5 |
| 2.2 to 2.3 | | | | | | | | 1 | 1 | 1 | 2 | | | | | | 5 |
| 2.1 to 2.2 | | | | | 1 | 1 | | 1 | 1 | 1 | 1 | | 1 | | 1 | | 8 |
| 2.0 to 2.1 | | | | 1 | 1 | | 2 | 1 | | 1 | | | | | | | 6 |
| 1.9 to 2.0 | | | | 1 | | 1 | 2 | 1 | 1 | 1 | | | | | | | 7 |
| 1.8 to 1.9 | | | | | | | | | | | | | | | | | 0 |
| 1.7 to 1.8 | | 1 | | | 1 | 1 | | | | | | | | | | | 3 |
| 1.6 to 1.7 | | | | | | | | | | | | | | | | | 0 |
| 1.5 to 1.6 | 1 | | | | | | | | | | | | | | | | 1 |
| 1.4 to 1.5 | | | | 1 | | | | | | | | | | | | | 1 |
| Totals | 1 | 1 | 0 | 3 | 3 | 3 | 4 | 4 | 3 | 5 | 3 | 2 | 4 | 2 | 2 | 1 | 41 |

The correlation found from the preceding table is $r = .789$ , a fairly high positive value. The usual formula for the probable error of a correlation coefficient derived from n pairs of observations, i.e., P.E. = $\frac{.6745(1-r^2)}{\sqrt{n}}$ would give for the above value of r a probable error of $\pm .041$ . It has, however, been pointed out by Fisher that the use of this formula is often misleading, especially in a case like this where the sample is small, and the correlation coefficient high. The distribution of correlation coefficients is often far from normal with small samples, and even with large samples if the correlation is high.

The method proposed as an alternative is to transform the coefficient by the substitution

$$z = \tfrac{1}{2}\log_e \frac{1+r}{1-r} \quad .$$

The distribution of z tends to normality rapidly as the sample is increased, whether r is large or not. The standard deviation of z is $\frac{1}{\sqrt{n-3}}$ , depending only on the number of observations. In Fisher's "Statistical Methods" there is provided a table (Table VB) giving the value of z for any r.

For $r = .789$ we find $z = 1.069$

$$\sigma_z = \frac{1}{\sqrt{38}} = .162$$

The value of z found therefore differs significantly (i.e. by $2\sigma$) from $1.069 \pm .324$ , i.e. from $z = 1.393$ and $z = .745$ . The corresponding values of r are .884 and .632 . There is thus no

doubt

dubiety about the significance of the correlation found; it is almost certainly greater than .632 , and less than .884 .

The regression equation showing the regression of $\sigma_N$ on $\sigma_{A.P.D.}$ is

$$\sigma_N = 0.55\,\sigma_{A.P.D.} + 0.34 ,$$

where $\sigma_N$ and $\sigma_{A.P.D.}$ are measured in the same units as used in their calculation originally. This estimate has a fairly high reliability by reason of the high value of r.

For a given value of $\sigma_{A.P.D.}$ the array of $\sigma_N$ has a standard deviation $s' = s\sqrt{1 - r^2}$ , where s is the standard deviation of the total array of all the $\sigma_N$ 's . In the present case, s was found to equal $\sqrt{2.68/41}$ so that

$$s' = \sqrt{\frac{2.68}{41}(1 - .789^2)} = .16$$

For example, if $\sigma_{A.P.D.} = 3$, the most probable value of $\sigma_N$ would be 1.99 , and about two thirds of the values of $\sigma_N$ expected would fall in the range 1.83 to 2.15 .

## 3. Significance of the Results.

These results are based on assumptions of normality in the arrays, assumptions whose truth can hardly be tested with so sm small an amount of data. It must also be remembered that they have been derived from the 41 tests and are strictly applicable only to those 41 tests. At the same time their general trend is significant.

Consider for example an answer-pattern of the "flat" type, as illustrated in the figure below.

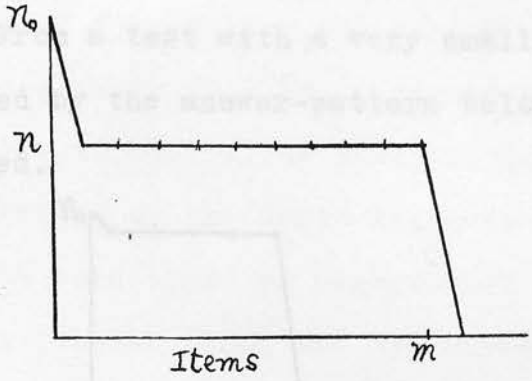



Figure 13. A flat answer-pattern.

The corresponding answer-pattern-differential is $n_o$-n, 0, 0, .....0, n. This has a large $\sigma_{A.P.D.}$ and therefore will probably produce a score-scatter with a large value of $\sigma_N$, that is the scores will be widely distributed about the mean.

Consider secondly a test of the "steep" type.

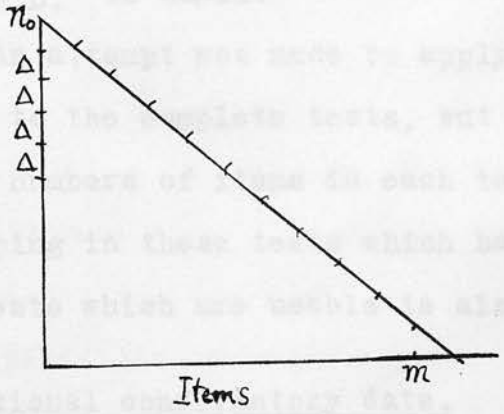



Figure 14. A steep answer-pattern.

This test has an answer-pattern-differential $\Delta, \Delta, \Delta, .., \Delta$, which has an intermediate value of $\sigma_{A.P.D.}$. The standard

deviation of the scaores in such a test would therefore be expected to be of average size for the number of items in the test.

From a test with a very small $\sigma_{A.P.D.}$, such as would be produced by the answer-pattern below, a very small $\sigma_{\bar{N}}$ would be expected.

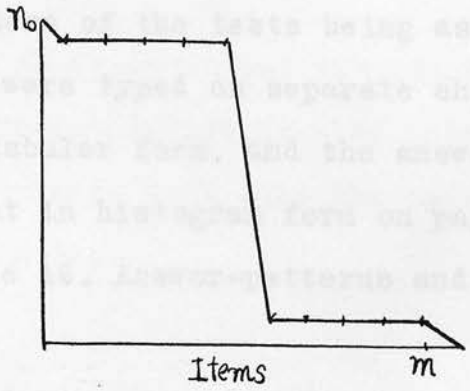



Figure 15. Answer-pattern producing small $\sigma_{A.P.D.}$.

Here the answer-pattern-differential is $n_o-n_1$, 0, 0,... ..0, $n_1-n_2$, 0, 0,....,$n_2$. If $n_o-n_1$ and $n_2$ are small, the value of $\sigma_{A.P.D.}$ is small.

An attempt was made to apply the above methods of investigation to the complete tests, but difficulties arose through the varying numbers of items in each test, and through the effects of grouping in these tests which had many items. The number of these tests which are usable is also small.

4. Additional confirmatory data.

As a practical test of these principles, three tests were designed to follow roughly the lines shown in figures 13-15. They were 8-item physics tests and were given to 34 pupils in the

first year of a secondary school.

Test D was designed as a fairly steep test of the type shown in figure 14 ; test E was a two level answer-pattern of the type of figure 15 ;,and test F was a rather flat test somewhat similar to that in figure 13. The usual precautions were taken to neutralise practice and fatigue effects, the various cyclic orders of the tests being as far as possible equally used. The tests were typed on separate sheets. The results are given below in tabular form, and the answer-patterns and score-scatters are aet out in histogram form on pages 85-87.

Table 16. Answer-patterns and score-scatters of tests D,E,F.

Test D

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Answer-pattern | 34, | 34, | 27, | 22, | 21, | 18, | 12, | 8, 5. |
| Score-scatter | 0, | 2, | 1, | 6, | 9, | 10, | 4, | 1, 1. |

| | |
|---|---|
| Standard deviation of A.P.D. | 2.48 |
| Standard deviation of score-scatter | 1.49 |
| Mean score | 4.3 |

Test E

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Answer-pattern | 34, | 32, | 30, | 30, | 29, | 9, | 7, | 6, 2. |
| Score-scatter | 0, | 0, | 4, | 2, | 17, | 6, | 3, | 1, 1. |

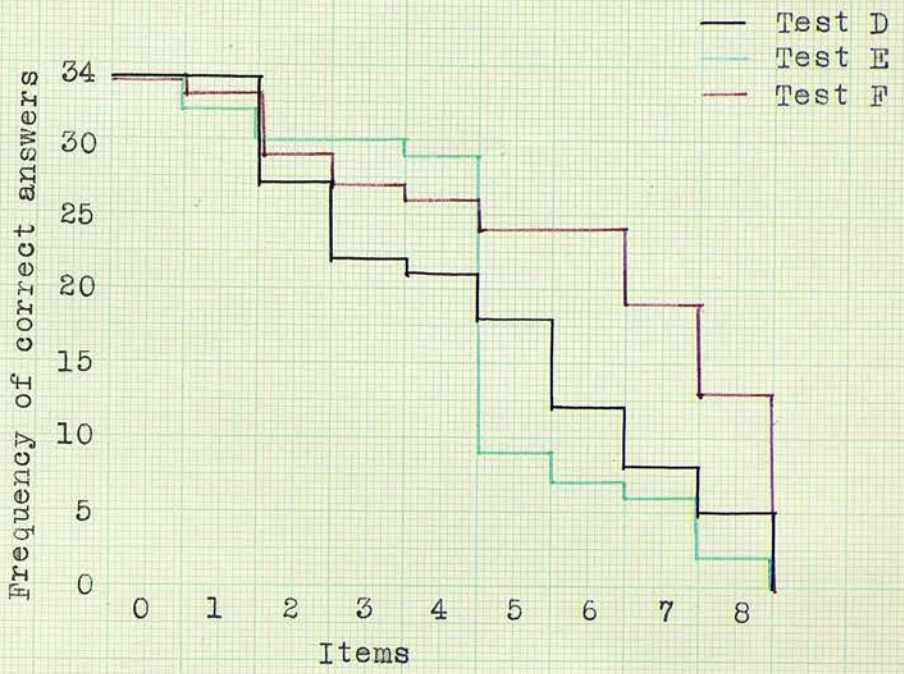
| | |
|---|---|
| Standard deviation of A.P.D. | 1.90 |
| Standard deviation of score-scatter | 1.31 |
| Mean score | 4.3 |

Test F

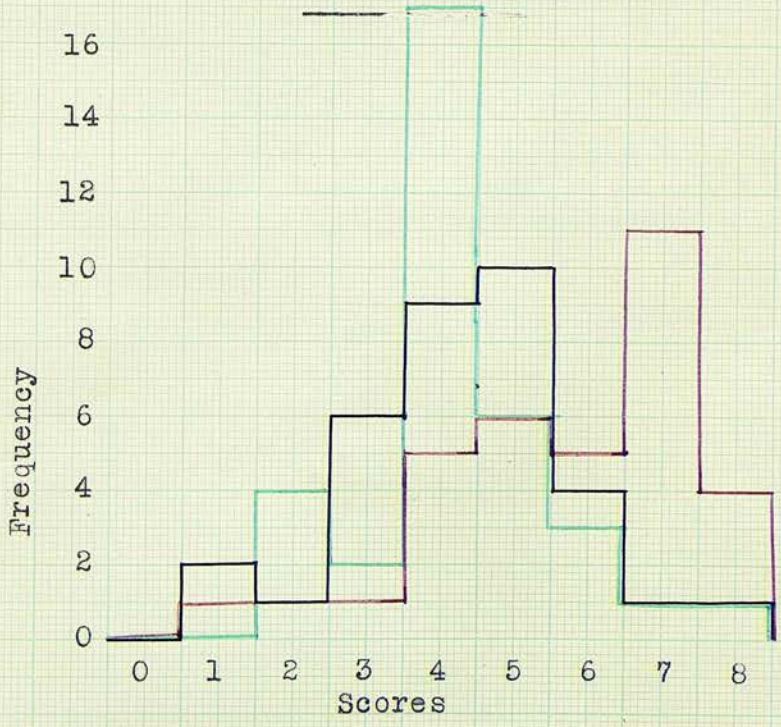
| | |
|---|---|
| Answer-pattern | 34, 33, 29, 27, 26, 24, 24, 19, 13. |
| Score-scatter | 0, 1, 1, 1, 5, 6, 5, 11, 4. |
| Standard deviation of A.P.D. | 2.66 |
| Standard deviation of score-scatter | 1.72 |
| Mean score | 5.7 |

It is obvious that the standard deviations of the scores vary as predicted. It is also interesting to note that the values fit quite well into the tail of the correlation table for the 10-item tests.

Fig. 16. Answer-patterns and score-scatters of Tests D, E, and F.




Answer-patterns.




Score-scatters.

## Chapter 7. The Correlation of the Coefficients of Skewness of Answer-pattern-differential and Score-scatter.

### 1. The measurement of skewness.

The skewness of a distribution such as a score-scatter is usually measured by the standardised value of the third moment about the mean. If the third moment $\frac{1}{n_o}\sum(x-a)^3$ is denoted by $\mu_3$, then the skewness S equals $\frac{\mu_3}{\sigma^3}$, where a, $n_o$, and $\sigma$ have their usual meanings.

The skewness so defined is positive when the curve has the tail to the right; negative when the curve has the tail to the left; and is zero for symmetrical distributions such as the normal frequency distribution.

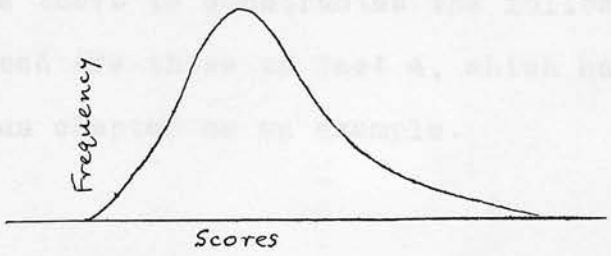



Figure 19. A positively skewed score-scatter.

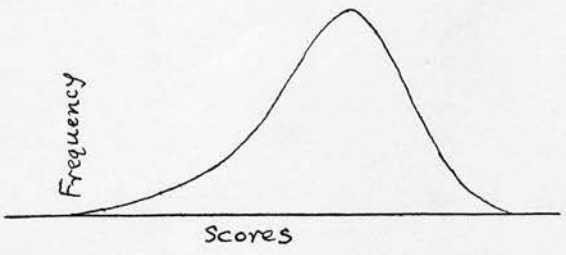



Figure 20. A negatively skewed score-scatter.

The application of this formula to the answer-pattern-differential may be justified in the same way as was done in the case of the standard deviations. The skewness of the answer-pattern-differential is denoted throughout this chapter by S'.

The method of calculation of S and S' follows closely that already shown for the standard deviations, except that the third moment must also be calculated. If the first, second, and third moments about the assumed mean are denoted by $m_1$, $m_2$, and $m_3$, then the skewness $\frac{\mu_3}{\sigma^3}$ can be shown to equal

$$\frac{m_3 - 3m_1m_2 + 2m_1^3}{(m_2 - m_1^2)^{3/2}}.$$

Thus there is constructed the following table, in which the data used are those of Test 4, which have already served in the previous chapter as an example.

Table 17. Calculation of S and S' for Test 4.

Answer-pattern-differential.

| Item | A.P. | A.P.D. | x | fx | $fx^2$ | $fx^3$ |
|---|---|---|---|---|---|---|
| 0 | 32 | 3 | -4 | -12 | 48 | -192 |
| 1 | 29 | 2 | -3 | - 6 | 18 | - 54 |
| 2 | 27 | 4 | -2 | - 8 | 16 | - 32 |
| 3 | 23 | 5 | -1 | - 5 | 5 | - 5 |
| 4 | 18 | 6 | 0 | 0 | 0 | 0 |
| 5 | 12 | 3 | 1 | 3 | 3 | 3 |
| 6 | 9 | 1 | 2 | 2 | 4 | 8 |
| 7 | 8 | 1 | 3 | 3 | 9 | 27 |
| 8 | 7 | 1 | 4 | 4 | 16 | 64 |
| 9 | 6 | 0 | 5 | 0 | 0 | 0 |
| 10 | 6 | 6 | 6 | 36 | 216 | 1296 |
| | 145 | $n_0$..32 | | +17 | +335 | +1115 |

$m_1 = 17/32 = +0.53$ ; $m_2 = 335/32 = +10.47$

$m_3 = 1115/32 = +34.84$

$$S' = \frac{m_3 - 3m_1m_2 + 2m_1^3}{(m_2 - m_1^2)^{3/2}} = +0.57$$

Score-scatter.

| Item | N | x | Nx | $Nx^2$ | $Nx^3$ |
|---|---|---|---|---|---|
| 0 | 1 | -4 | -4 | 16 | -64 |
| 1 | 1 | -3 | -3 | 9 | -27 |
| 2 | 1 | -2 | -2 | 4 | - 8 |
| 3 | 10 | -1 | -10 | 10 | -10 |
| 4 | 5 | 0 | 0 | 0 | 0 |
| 5 | 4 | 1 | 4 | 4 | 4 |
| 6 | 4 | 2 | 8 | 16 | 32 |
| 7 | 2 | 3 | 6 | 18 | 54 |
| 8 | 2 | 4 | 8 | 32 | 128 |
| 9 | 2 | 5 | 10 | 50 | 250 |
| 10 | 0 | 6 | 0 | 0 | 0 |
| | $n_0$..32 | | +17 | +159 | +359 |

$m_1 = +0.53$ $m_2 = +4.98$ $m_3 = +11.24$

$S = +0.36$

2. Results of the first experiment.

In this way the skewness of the answer-pattern-differential (S') and of the score-scatter (S) were calculated for each of the 41 tests. The results were as follows.

Table 18. Coefficients of skewness of the 41 tests.

| Test | S' | S |
|---|---|---|
| 1 | -0.015 | -0.28 |
| 2 | -0.09 | -0.17 |
| 3 | -0.38 | +0.15 |
| 4 | +0.57 | +0.36 |
| 5 | -0.52 | -0.003 |
| 6 | -0.19 | +0.33 |
| 7 | -0.09 | -0.01 |
| 8 | -0.76 | -0.32 |
| 9 | +0.24 | -0.06 |
| 10 | -0.603 | -0.42 |
| 11 | -1.08 | -0.78 |
| 12 | -0.64 | -1.12 |
| 13 | -0.42 | -0.23 |
| 14 | -0.35 | -0.59 |
| 15 | +0.14 | +0.103 |
| 16 | +0.206 | +0.02 |
| 17 | -1.002 | -0.55 |
| 18 | -0.23 | -0.14 |
| 19 | -0.41 | +0.04 |
| 20 | -0.05 | +0.21 |
| 21 | -1.01 | -0.297 |
| 22 | -0.15 | +0.12 |
| 23 | -0.92 | -0.93 |
| 24 | -0.69 | -0.28 |
| 25 | -0.93 | -0.74 |
| 26 | +0.99 | -0.74 |
| 27 | -0.799 | -0.27 |
| 28 | -0.59 | -0.202 |
| 29 | -0.52 | -0.78 |
| 30 | -0.55 | -0.23 |
| 31 | -1.06 | -0.95 |
| 32 | -0.95 | +0.07 |
| 33 | -1.31 | -0.38 |
| 34 | -0.75 | +0.004 |
| 35 | -0.53 | -0.13 |
| 36 | -0.53 | -0.35 |
| 37 | -0.55 | -0.28 |
| 38 | -0.43 | -0.13 |
| 39 | -0.42 | -0.404 |
| 40 | -0.94 | -0.64 |
| 41 | -0.48 | -0.69 |

The third decimal place has been added where required for grouping in the correlation table. This table is shown on page 91.

The correlation between S and S' was found to be $r = .628$. The probable error which might be attached to this coefficient is $\pm$ 0.064, or, using the transformation $z = \frac{1}{2}\log_e \frac{1+r}{1-r}$, we find that r differs significantly from .738 and .392. The correlation is not so high as that obtained with $\sigma_{A.P.D.}$ and $\sigma_N$ but it still is quite significant.

The regression equation is

$$S = 0.55S' + 0.00$$

and the standard deviation of an array of S is 0.28. The units used are those in which S and S' were originally measured.

3. The natural skewness of a population.

A rather interesting point may be dealt with here. Let us suppose that the population tested has a "natural" skewness, as opposed to the skewness forced on it by the answer-pattern. A possible way of measuring this natural skewness is to apply a test for which S' = 0; or, better still, to apply a battery of tests such as the 41 tests, and evaluate S from the regression equation obtained, substituting S' = 0 in that equation.

From the equation $S = 0.55S' + 0.00$ we find that the skewness of the population tested by the 41 tests is zero. This estimate has of course a standard deviation of 0.28, as seen above.

Table 19. Correlation of coefficients of skewness of answer-pattern-differential and score-scatter (41 tests). $r = .628$.

| S' \ S | -1.2 to -1.1 | -1.1 to -1.0 | -1.0 to -0.9 | -0.9 to -0.8 | -0.8 to -0.7 | -0.7 to -0.6 | -0.6 to -0.5 | -0.5 to -0.4 | -0.4 to -0.3 | -0.3 to -0.2 | -0.2 to -0.1 | -0.1 to -0.0 | 0.0 to 0.1 | 0.1 to 0.2 | 0.2 to 0.3 | 0.3 to 0.4 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 to 0.6 | | | | | | | | | | | | | | | | 1 | 1 |
| 0.4 0.5 | | | | | | | | | | | | | | | | | 0 |
| 0.3 0.4 | | | | | | | | | | | | | | | | | 0 |
| 0.2 0.3 | | | | | | | | | | | | 1 | 1 | | | | 2 |
| 0.1 0.2 | | | | | | | | | | | | | | 1 | | | 1 |
| 0.0 0.1 | | | | | | | | | | | | | | | | | 0 |
| -0.1 0.0 | | | | | | | | | | 1 | 1 | 1 | | | 1 | | 4 |
| -0.2 -0.1 | | | | | | | | | | | | | | 1 | | 1 | 2 |
| -0.3 -0.2 | | | | | | | | | | | 1 | | | | | | 1 |
| -0.4 -0.3 | | | | | | | 1 | | | | | | | 1 | | | 2 |
| -0.5 -0.4 | | | | | | 1 | | 1 | | 1 | 1 | | 1 | | | | 5 |
| -0.6 -0.5 | | | | | 1 | | | | 1 | 3 | 1 | 1 | | | | | 7 |
| -0.7 -0.6 | 1 | | | | | | | 1 | | 1 | | | | | | | 3 |
| -0.8 -0.7 | | | | | | | | | 1 | 1 | | | 1 | | | | 3 |
| -0.9 -0.8 | | | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | 1 | | 2 | 1 | | | | | | | 1 | | | | 5 |
| -1.1 -1.0 | | | 1 | | 1 | | 1 | | | 1 | | | | | | | 4 |
| -1.2 -1.1 | | | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | | | | | | | | | | | | | | | | | 0 |
| -1.4 -1.3 | | | | | | | | | 1 | | | | | | | | 1 |
| Totals | 1 | 0 | 2 | 0 | 4 | 2 | 2 | 2 | 3 | 8 | 4 | 3 | 4 | 3 | 1 | 2 | 41 |

4. Second experiment.

We have reason to believe that mental ability like so many physical characteristics is distributed normally. Now the skewness of a normal distribution is zero. If the population tested is sufficiently large and unselected, it seems that we may assume its natural skewness to be zero, and any skewness found in the score-scatters to be attributable to the effect of a skewed answer-pattern-differential.

This method was applied to the complete tests already mentioned. The numbers tested are sufficiently large to obviate large deviations from the natural skewness zero, The fact that different candidates are under examination in each test does not affect the results.

In some cases the application of the method of calculation already shown was quite straightforward; in other cases, such as test M.H.T. 8, which has 109 items, the score-scatter and answer-pattern-differential had to be grouped in the usual way before evaluation of S and S'. There were 22 tests in all. The results are shown overleaf.

Table 20. Coefficients of skewness of complete tests.

| Test | S' | S |
|---|---|---|
| M.H.T. 8 | -0.38 | -0.66 |
| M.H.T. 9 | +0.27 | -0.06 |
| M.H.T. 11 | +0.15 | -0.15 |
| M.H.T. 12v | +0.199 | -0.17 |
| M.H.T. 12p | -1.201 | -1.33 |
| Thesis B | +0.23 | +0.01 |
| Thesis C | -0.28 | -0.24 |
| A II | -0.16 | -0.19 |
| A IX | -0.13 | -0.31 |
| A X | -0.36 | -0.32 |
| A XI | -0.28 | -0.38 |
| A XIII | -0.43 | -0.22 |
| A XV | +0.19 | -0.08 |
| A XXVI | -0.13 | -0.39 |
| A XXXIV | -0.26 | +0.07 |
| A XXXII | -0.12 | -0.15 |
| A XXXIII | +0.14 | -0.08 |
| K | +0.14 | +0.09 |
| L | -0.21 | +0.01 |
| K2 | -0.11 | -0.36 |
| M | -0.39 | -0.53 |
| M2 | -0.48 | -0.31 |

The correlation table showing the relation between S and S' is shown on page 94. The correlation was found to be r = .836 Proceeding as before we find that this coefficient, although derived from a comparatively small number of cases, is almost certainly greater than .627 and less than .929 .

The regression equation is

$$S = 0.77S' - 0.13 ,$$

with a standard deviation in the arrays of S of 0.17 .

Substituting S' = 0, we find that the natural skewness of the population tested is -0.13 , which does not differ significantly from zero, since the standard deviation of any array is 0.17 .

Table 21. Coefficients of skewness of answer-pattern-differential and score-scatter in correlation table form. (22 complete tests). $r = .836$.

| S' | S: -1.4 to -1.3 | -1.3 to -1.2 | -1.2 to -1.1 | -1.1 to -1.0 | -1.0 to -0.9 | -0.9 to -0.8 | -0.8 to -0.7 | -0.7 to -0.6 | -0.6 to -0.5 | -0.5 to -0.4 | -0.4 to -0.3 | -0.3 to -0.2 | -0.2 to -0.1 | -0.1 to 0.0 | 0.0 to 0.1 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 to 0.3 | | | | | | | | | | | | | | 1 | 1 | 2 |
| 0.1 0.2 | | | | | | | | | | | | | 2 | 2 | 1 | 5 |
| 0.0 0.1 | | | | | | | | | | | | | | | | 0 |
| -0.1 0.0 | | | | | | | | | | | | | | | | 0 |
| -0.2 -0.1 | | | | | | | | | | | 3 | | 2 | | | 5 |
| -0.3 -0.2 | | | | | | | | | | | 1 | 1 | | | 2 | 4 |
| -0.4 -0.3 | | | | | | | | 1 | 1 | | 1 | | | | | 3 |
| -0.5 -0.4 | | | | | | | | | | | 1 | 1 | | | | 2 |
| -0.6 -0.5 | | | | | | | | | | | | | | | | 0 |
| -0.7 -0.6 | | | | | | | | | | | | | | | | 0 |
| -0.8 -0.7 | | | | | | | | | | | | | | | | 0 |
| -0.9 -0.8 | | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | | | | | | | | | | | | | | 0 |
| -1.1 -1.0 | | | | | | | | | | | | | | | | 0 |
| -1.2 -1.1 | | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | 1 | | | | | | | | | | | | | | | 1 |
| [illegible] | | | | | | | | | | | | | | | | |
| Totals | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 6 | 2 | 4 | 3 | 4 | 22 |

5. Significance of the experimental results.

It follows from the positive correlation between S and S' shown in the results of both these experiments that a positively skewed score-scatter is more likely to occur with a positively skewed answer-pattern-differential, and the extent to which it is skewed will depend partly on the degree of skewness of that distribution. To construct a test which is intended to give a score-scatter skewed positively, the examiner should therefore work with an answer-pattern falling steeply at first and then flattening out as it nears the items axis. Conversely, a test designed to produce a negatively skewed score-scatter should have an answer-pattern falling gently at first, and increasing in slope to a maximum. Then the nature of the correlation between S and S' found in the above experiments shows that the skewing of the answer-pattern-differential so caused will be accompanied by a similar skewing of the score-scatter to an extent indicated by the size of the correlation coefficient.

6. The influence of difficulty level on skewness of score-scatter

It is a fact already well known to examiners that score-scatters may be skewed by suitable adjustment of the difficulty level of a test. A recently published book on the science of marking says,

"If the curve is skewed to the low side, it means that marks have been difficult to get, which may be due.....to the questions

being too difficult. On the other hand, a curve skewed toward the upper part of the mark scale suggests that the paper has been easy. "

The problem that at once arises is the relation of this fact to the above theory. In the subsequent discussion I hope to make it clear that the use of the difficulty level as a method of skewing score-scatters is, in fact, merely an approximation to the use of the answer-pattern-differential.

The relation between difficulty level and skewness of score-scatter may be studied in the same way as the relation between S and S', provided we devise some measure of difficulty. This is comparatively easy. If the ratio of the total number of correct answers in a test to the total possible number is e, then the difficulty level d may be defined as 1 - e. In the notation used in previous chapters, $d = 1 - \frac{\Sigma n}{mn_o}$ .

The difficulties of all the tests used were calculated and are tabulated overleaf.

Table 22. Difficulty levels of tests.

(1) The 41 tests.

| Test | d | Test | d |
|---|---|---|---|
| 1 | .57 | 22 | .47 |
| 2 | .44 | 23 | .27 |
| 3 | .500 | 24 | .31 |
| 4 | .55 | 25 | .25 |
| 5 | .36 | 26 | .25 |
| 6 | .42 | 27 | .35 |
| 7 | .42 | 28 | .37 |
| 8 | .31 | 29 | .39 |
| 9 | .503 | 30 | .33 |
| 10 | .31 | 31 | .303 |
| 11 | .28 | 32 | .34 |
| 12 | .31 | 33 | .19 |
| 13 | .42 | 34 | .36 |
| 14 | .46 | 35 | .33 |
| 15 | .51 | 36 | .33 |
| 16 | .41 | 37 | .32 |
| 17 | .28 | 38 | .38 |
| 18 | .49 | 39 | .397 |
| 19 | .41 | 40 | .29 |
| 20 | .43 | 41 | .45 |
| 21 | .29 | | |

(2) The complete tests.

| Test | d | Test | d |
|---|---|---|---|
| M.H.T. 8 | .33 | A XIII | .44 |
| M.H.T. 9 | .56 | A XV | .56 |
| M.H.T. 11 | .505 | A XXVI | .51 |
| M.H.T. 12v | .52 | A XXXIV | .49 |
| M.H.T. 12p | .25 | A XXXII | .59 |
| Thesis B | .66 | A XXXIII | .58 |
| Thesis C | .53 | K | .52 |
| A II | .46 | L | .43 |
| A IX | .38 | K2 | .45 |
| A X | .39 | M | .42 |
| A XI | .41 | M2 | .46 |

Proceeding to the calculation of the correlation between the skewness (S) of the score-scatter and the difficulty level (d) of the answer-pattern, we find that for the 41 tests $r_{Sd}$ equals .561. and for the 22 tests equals .790 . These compare with $r_{SS'}$ = .628 for the 41 tests and .836 for the 22 tests. The correlation tables are shown on pages 99 and 100.

It will be observed that in each case the correlation between S and S' is greater than the correlation between S and d. On the other hand it must be pointed out that the difference is not statistically significant. To test the significance of the difference, it is necessary to use the transformation $z = \frac{1}{2}\log_e \frac{1+r}{1-r}$ For the 41 tests $r_{SS'}$ = .628, $r_{Sd}$ = .561; the corresponding values of z are .738 and .634, giving a difference .104. The standard deviation of this difference equals the square root of the sum of the squares of the standard deviations of the z's, and these deviations, as indicated in chapter 6, are each $\frac{1}{\sqrt{38}}$. The standard deviation of the difference is therefore $\sqrt{\frac{1}{38} + \frac{1}{38}}$ = .229. Thus the difference is less than its standard error, and so is not significant. The lack of significance is even more strongly shown in the complete tests, their fewness increasing the standard deviation of the difference.

On the other hand, of all the various groups of tests which where afterwards selected from the whole 63 tests for various purposes, none has been found with $r_{SS'}$ less than $r_{Sd}$ .

Table 23. Correlation of S and d , 41 tests. $r = .561$.

| S | d: .18 to .21 | .21 to .24 | .24 to .27 | .27 to .30 | .30 to .33 | .33 to .36 | .36 to .39 | .39 to .42 | .42 to .45 | .45 to .48 | .48 to .51 | .51 to .54 | .54 to .57 | .57 to .60 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.3 to 0.4 | | | | | | | | | 1 | | | | 1 | | 2 |
| 0.2 0.3 | | | | | | | | | 1 | | | | | | 1 |
| 0.1 0.2 | | | | | | | | | | 1 | 1 | 1 | | | 3 |
| 0.0 0.1 | | | | | | 1 | 1 | 2 | | | | | | | 4 |
| -0.1 0.0 | | | | | | | 1 | | 1 | | 1 | | | | 3 |
| -0.2 -0.1 | | | | | | 1 | 1 | | 1 | | 1 | | | | 4 |
| -0.3 -0.2 | | | | 1 | 2 | 2 | 1 | | 1 | | | | | 1 | 8 |
| -0.4 -0.3 | 1 | | | | 1 | 1 | | | | | | | | | 3 |
| -0.5 -0.4 | | | | | 1 | | | 1 | | | | | | | 2 |
| -0.6 -0.5 | | | | 1 | | | | | | 1 | | | | | 2 |
| -0.7 -0.6 | | | | 1 | | | | | | 1 | | | | | 2 |
| -0.8 -0.7 | | | 2 | 1 | | | | 1 | | | | | | | 4 |
| -0.9 -0.8 | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | | 1 | 1 | | | | | | | | | | 2 |
| -1.1 -1.0 | | | | | | | | | | | | | | | 0 |
| -1.2 -1.1 | | | | | 1 | | | | | | | | | | 1 |
| Totals | 1 | 0 | 2 | 5 | 6 | 5 | 4 | 4 | 5 | 3 | 3 | 1 | 1 | 1 | 41 |

Table 24. Correlation of S and d, 22 tests. $r = .790$.

| S | d .24 to .27 | .27 to .30 | .30 to .33 | .33 to .36 | .36 to .39 | .39 to .42 | .42 to .45 | .45 to .48 | .48 to .51 | .51 to .54 | .54 to .57 | .57 to .60 | .60 to .63 | .63 to .66 | .66 to .69 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 to 0.1 | | | | | | | 1 | | 1 | 1 | | | | | 1 | 4 |
| -0.1 0.0 | | | | | | | | | | | 2 | 1 | | | | 3 |
| -0.2 -0.1 | | | | | | | | 1 | 1 | 1 | | 1 | | | | 4 |
| -0.3 -0.2 | | | | | | | 1 | | | 1 | | | | | | 2 |
| -0.4 -0.3 | | | | | 1 | 2 | 1 | 1 | 1 | | | | | | | 6 |
| -0.5 -0.4 | | | | | | | | | | | | | | | | 0 |
| -0.6 -0.5 | | | | | | | 1 | | | | | | | | | 1 |
| -0.7 -0.6 | | | | 1 | | | | | | | | | | | | 1 |
| -0.8 -0.7 | | | | | | | | | | | | | | | | 0 |
| -0.9 -0.8 | | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | | | | | | | | | | | | | | 0 |
| -1.1 -1.0 | | | | | | | | | | | | | | | | 0 |
| -1.2 -1.1 | | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | | | | | | | | | | | | | | | | 0 |
| -1.4 -1.3 | 1 | | | | | | | | | | | | | | | 1 |
| Totals | 1 | 0 | 0 | 1 | 1 | 2 | 4 | 2 | 3 | 3 | 2 | 2 | 0 | 0 | 1 | 22 |

7. The linearity of regression of S on various functions of d.

In the correlation of S with d , there is a difficulty which did not arise in previous correlations, the difficulty of linearity of regression. S and S' are both of the third order of moments, while d is of the first order. From this point of view it would seem that we should correlate S and $d^3$ instead of S and d. To avoid errors due to the introduction of non-linear regressions, all the correlations given above were tested for linearity, as also were the regression of S' on d, of S on $d^3$, and of S' on $d^3$ for both sets of data.

The method of testing the linearity of regression was that set out in Fisher's"Statistical Methods". This is a comparatively new method, replacing the older method using Blakeman's criterion, which was used in the author's second paper in the British Journal of Psychology. The objections to the validity of the older method are to be found in Fisher's book, section 46. In an appendix to this chapter there is shown in full the application of the new method to one of the above correlations.

For an understanding of the significance of the results it is necessary to give a very brief explanation of the method. Consider the correlation of S with d. For each value of d there is an array of values of S. The regression line of S on d is the straight line of best fit to the means of these arrays. The deviations of the means from this line represent the departures from linearity of regression; these deviations may be compared

with the deviations of the single observations from the mean of their array. Whether the deviations from linearity are significantly greater than those in the arrays is determined by calculating a variable z; then the probabilty of a given value of z being exceeded through chance variations is tabulated in Table VI of "Statistical Methods". The 5% point , representing a probabilty of 1 in 20 is usually taken as the dividing line.

When this test was applied to the various correlations, the results were in some ways surprising. The correlations $r_{SS'}$, $r_{Sd}$, $r_{S'd}$ for both groups of tests were sufficiently linear, with the exception of $r_{SS'}$ for the 22 tests. The value of z in this case was .70 while the 5% point was .53 and the 1% point .76. The deviations found would be exceeded only once in a hundred trials if the regression were linear. The lack of linearity is not due to any curving of the line of means; an examination of the correlation table shows it to be due rather to the zigzag nature of that line. It is probably caused by the mixing of data from various populations. An interesting point is that the correlation $r_{SS'}$ calculated from the 41 tests showed deviations from linearity less than the deviations within the arrays.

When the test was applied to the correlations involving $d^3$ it was found that all but one of these were also linear, the exception being the regression of S on $d^3$ for the 22 tests, where the value of z was .554 with a 5% point .548, so that the regressio is barely linear. Thus we have the apparent anomaly that S is

correlated in linear fashion with both d and $d^3$. This apparent anomaly will be cleared up later. Meanwhile, the test of linearity of regression applied to the data available does not enable us to distinguish between d and $d^3$ as to suitability for correlating with S.

8. The regression equation predicting S from S' and d.

Since S is found to be correlated both with S' and d, an obvious step is to construct the regression equation predicting S from S' and d. This equation is given by the formula

$$S = \frac{\Delta_{SS'}}{\Delta_{SS}} S' - \frac{\Delta_{Sd}}{\Delta_{SS}} d$$

where S, S', and d are measured in $\sigma$ units from their means and $\Delta_{SS'}$..... are the cofactors of $r_{SS'}$..... in the determinant

$$\Delta = \begin{vmatrix} r_{SS} & r_{SS'} & r_{Sd} \\ r_{SS'} & r_{S'S'} & r_{S'd} \\ r_{Sd} & r_{S'd} & r_{dd} \end{vmatrix}$$

The multiple correlation, R, of S with the team of S' and d weighted as above, is equal to $\sqrt{1 - \frac{\Delta}{\Delta_{SS}}}$

Before this equation can be set up, it is necessary to calculate the correlation of S' and d; this is readily obtainable from the data already given. From the 41 tests we find $r_{S'd}$ equal to .869 , and from the 22 tests $r_{S'd}$ equal to .805 . The tables are shown on pages 104 and 105.

Table 25. Correlation of S' and d, 41 tests.

| S' | d: .18 to .21 | .21 to .24 | .24 to .27 | .27 to .30 | .30 to .33 | .33 to .36 | .36 to .39 | .39 to .42 | .42 to .45 | .45 to .48 | .48 to .51 | .51 to .54 | .54 to .57 | .57 to .60 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 to 0.6 | | | | | | | | | | | | | 1 | | 1 |
| 0.4 0.5 | | | | | | | | | | | | | | | 0 |
| 0.3 0.4 | | | | | | | | | | | | | | | 0 |
| 0.2 0.3 | | | | | | | | 1 | | | 1 | | | | 2 |
| 0.1 0.2 | | | | | | | | | | | | 1 | | | 1 |
| 0.0 0.1 | | | | | | | | | | | | | | | 0 |
| -0.1 0.0 | | | | | | | | | 3 | | | | | 1 | 4 |
| -0.2 -0.1 | | | | | | | | | 1 | 1 | | | | | 2 |
| -0.3 -0.2 | | | | | | | | | | | 1 | | | | 1 |
| -0.4 -0.3 | | | | | | | | | | 1 | 1 | | | | 2 |
| -0.5 -0.4 | | | | | | | 1 | 2 | 1 | 1 | | | | | 5 |
| -0.6 -0.5 | | | | | 1 | 3 | 2 | 1 | | | | | | | 7 |
| -0.7 -0.6 | | | | | 3 | | | | | | | | | | 3 |
| -0.8 -0.7 | | | | | 1 | 1 | 1 | | | | | | | | 3 |
| -0.9 -0.8 | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | 2 | 2 | | 1 | | | | | | | | | 5 |
| -1.1 -1.0 | | | | 3 | 1 | | | | | | | | | | 4 |
| -1.2 -1.1 | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | | | | | | | | | | | | | | | 0 |
| -1.4 -1.3 | 1 | | | | | | | | | | | | | | 1 |
| Totals | 1 | 0 | 2 | 5 | 6 | 5 | 4 | 4 | 5 | 3 | 3 | 1 | 1 | 1 | 41 |

Table 26. Correlation of S' and d, 22 tests.

| S' | d: .24 to .27 | .27 to .30 | .30 to .33 | .33 to .36 | .36 to .39 | .39 to .42 | .42 to .45 | .45 to .48 | .48 to .51 | .51 to .54 | .54 to .57 | .57 to .60 | .60 to .63 | .63 to .66 | .66 to .69 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 to 0.3 | | | | | | | | | | | 1 | | | | 1 | 2 |
| 0.1 0.2 | | | | | | | | | 1 | 2 | 1 | 1 | | | | 5 |
| 0.0 0.1 | | | | | | | | | | | | | | | | 0 |
| -0.1 0.0 | | | | | | | | | | | | | | | | 0 |
| -0.2 -0.1 | | | | | 1 | | 1 | 1 | 1 | | | 1 | | | | 5 |
| -0.3 -0.2 | | | | | | 1 | 1 | | 1 | 1 | | | | | | 4 |
| -0.4 -0.3 | | | | 1 | | 1 | 1 | | | | | | | | | 3 |
| -0.5 -0.4 | | | | | | | 1 | 1 | | | | | | | | 2 |
| -0.6 -0.5 | | | | | | | | | | | | | | | | 0 |
| -0.7 -0.6 | | | | | | | | | | | | | | | | 0 |
| -0.8 -0.7 | | | | | | | | | | | | | | | | 0 |
| -0.9 -0.8 | | | | | | | | | | | | | | | | 0 |
| -1.0 -0.9 | | | | | | | | | | | | | | | | 0 |
| -1.1 -1.0 | | | | | | | | | | | | | | | | 0 |
| -1.2 -1.1 | | | | | | | | | | | | | | | | 0 |
| -1.3 -1.2 | 1 | | | | | | | | | | | | | | | 1 |
| Totals | 1 | 0 | 0 | 1 | 1 | 2 | 4 | 2 | 3 | 3 | 2 | 2 | 0 | 0 | 1 | 22 |

The data from the 41 tests are thus

$r_{SS'} = .628$, $r_{Sd} = .561$, $r_{S'd} = .869$,

and the resulting equation is $S = 0.57\ S' + 0.06\ d$.

The multiple correlation R is .629 .

Similarly from the 22 tests with correlations

$r_{SS'} = .836$, $r_{Sd} = .790$, $r_{S'd} = .805$,

there is obtained the equation $S = 0.57\ S' + 0.33\ d$, and R equals .876.

The first regression equation states that S' is 0.57/0.06 times as important as d in predicting S, and that the correlation of S with the team of S' and d, weighted in the best possible way, is no greater than the correlation of S with S'.

In the case of the complete tests the result is not quite so definite. S' is about twice as important as d in predicting S, and the correlation of S with the team of S' and d is only slightly greater than the correlation of S with S' alone.

The most significant thing in the calculation is the high value of $r_{S'd}$. The large size of this correlation coefficient is the main cause of the team of S' and d being little superior to S' alone as a criterion of S. It seems that in measuring S' and d, we are, to a great extent, measuring the same thing. This, in fact, is the key to the whole problem.

9. The relation of S' and d.

The high correlation between S' and d might have been expected for the following reasons.

In an answer-pattern graph such as that shown below, the difficulty d of the test with answer-pattern (a) may be represented by the shaded area, for the area between the curve, OP, and OQ, is a measure of the total number of correct answers.

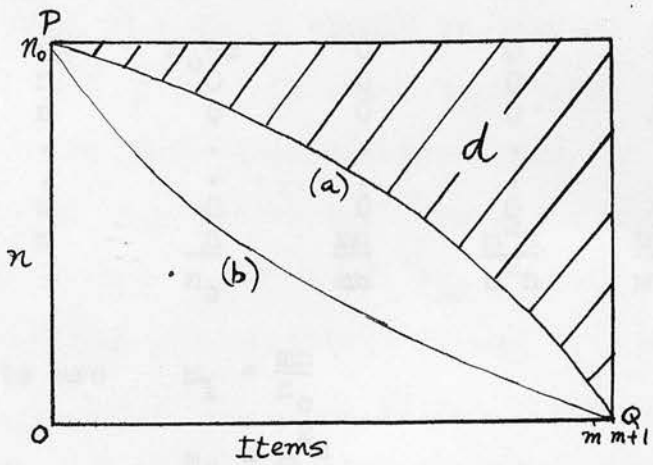



Figure 21. Answer-patterns of easy and difficult tests.

Now the answer-pattern of an easy test such as (a) will obviously produce a negatively skewed answer-pattern-differential On the other hand, a more difficult test, such as (b) will have a positively skewed answer-pattern-differential. It is possible to construct answer-patterns to which these arguments do not apply, but they are of peculiar type, such as tests half of whose items are answered correctly by all the candidates. These freak tests need not be considered here. In the ordinary test we may expect difficulty level and skewness of answer-pattern-differential

to be positively correlated.

In the case of a flat test, the skewness of the answer-pattern-differential is directly calculable from the difficulty level. Consider a test of m items, sat by $n_o$ candidates, each item being answered correctly n times. The answer-pattern is then $n_o$, n, n, .....n, n. The calculation of the skewness proceeds as before (cf. page 88)

| Item (x) | A.P. | A.P.D. | fx | $fx^2$ | $fx^3$ |
|---|---|---|---|---|---|
| 0 | $n_o$ | $n_o-n$ | 0 | 0 | 0 |
| 1 | n | 0 | 0 | 0 | 0 |
| 2 | n | 0 | 0 | 0 | 0 |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| m-1 | n | 0 | 0 | 0 | 0 |
| m | n | n | mn | $m^2n$ | $m^3n$ |
| | | $n_o$ | mn | $m^2n$ | $m^3n$ |

The moments are $m_1 = \frac{mn}{n_o}$

$$m_2 = \frac{m^2n}{n_o}$$

$$m_3 = \frac{m^3n}{n_o}$$

Substituting these values in the formula

$$S' = \frac{m_3 - 3m_1m_2 + 2m_1^3}{(m_2 - m_1^2)^{3/2}}$$

and using the identities

$$d = 1 - \frac{mn}{mn_o} = 1 - \frac{n}{n_o}$$

we find

$$S' = \frac{2d-1}{\sqrt{d(1-d)}}$$

This function of d may be denoted by $\phi(d)$. For any flat test the skewness of the answer-pattern-differential is therefore directly calculable from the difficulty level. Consider, for example, a test in which each item is answered correctly by half the candidates. Here $n = n_o/2$ and therefore $d = 0.5$. Substituting in the formula we find $S' = 0$. The answer-pattern-differential of such a test is therefore unskewed.

In passing, we may note that the formula suggests that the function of d which should be correlated with S is neither d nor $d^3$ but $\phi(d)$. To check this point, $\phi(d)$ was evaluated for the tests under consideration, and the linearity of the correlations $r_{s\phi}$ was tested. From the 41 tests the value of $r_{s\phi}$ obtained was .576 ( cf. $r_{Sd} = .561$, $r_{Sd^3} = .490$ ), and the correlation was certainly linear ( z = .15, 5% point .38 ). From the 22 tests the value of $r_{s\phi}$ obtained was .816 ( cf. $r_{Sd} = .790$, $r_{Sd^3} = .620$ ) and the correlation was again linear, ( z = .34, 5% point .55 ). The increase in the size of the correlation coefficient points to the employment of a better statistical method.

Where the test is not perfectly flat, the skewness of the answer-pattern-differential is not equal to $\phi(d)$. But it can be demonstrated that in any test $\phi$ is a good approximation to S', and d is a good approximation to $\phi$. The close relation of S' and d is a corollary of the theorem that d is an approximation, through $\phi$, to S'.

It will readily be seen that the first approximation, of $\phi$ for S', consists in replacing the answer-pattern of the test, from which S' is calculable, by the answer-pattern of a flat test of the same difficulty level. Instead of using all the distinctive values of n which make up the answer-pattern, the tester uses only the average of these values. Symbolically, the process is as follows.

The skewness S' of the answer-pattern-differential is given by the formula

$$n_0\sigma^3 S' = \sum_{x=0}^{m}(n_x - n_{x+1})(x - a)^3$$

where a is the mean score.

Let n be the mean value of $n_1, n_2, \ldots\ldots\ldots, n_m$,

and define $\nu_x$ as $n_x - n$.

Then $$n_x - n_{x+1} = \nu_x - \nu_{x+1}$$

and $$n_m = \nu_m + n$$

$$n_0\sigma^3 S' = (\nu_0 - \nu_1)(0-a)^3 + (\nu_1 - \nu_2)(1-a)^3 + \ldots\ldots\ldots$$

$$\ldots + (\nu_{m-1} - \nu_m)(m-1-a)^3 + (\nu_m + n)(m-a)^3.$$

$$= -\nu_0 a^3 + \nu_1(3a^2-3a+1) + \ldots\ldots + \nu_m\left[3(m-a)^2-3(m-a)+1\right] + n(m-a)^3.$$

Now $\nu_0 > \nu_1 \geqslant \nu_2 \geqslant \nu_3 \geqslant \ldots\ldots$ as far as $\nu_x$ is positive,

and $a^3 > 3a^2-3a+1 > \ldots\ldots\ldots$ since $a > 1$.

Therefore the first term is greater than the second, and so on.

Also $n > |\nu_m| \geqslant |\nu_{m-1}| \geqslant \ldots\ldots\ldots$ where $\nu_x$ is negative,

and $(m-a)^3 \geqslant 3(m-a)^2-3(m-a)+1 \geqslant \ldots$ since $m > a+1$.

Therefore the last term is greater than the second last, and so on.

A first approximation to S' is therefore given by using only the first and last terms of the series.

$$\therefore n_o \sigma^3 S' = -\gamma_o a^3 + n(m-a)^3.$$

Dividing through by $n_o m^3$, and using the identities

$$d = 1 - \frac{a}{m} = 1 - \frac{n}{n_o},$$

we have $$\frac{\sigma^3}{m^3} S' = -d(1-d)^3 + (1-d)d^3$$

i.e. $$S' = \frac{m^3}{\sigma^3} \left[d(1-d)(2d-1)\right]$$

This is precisely the formula previously obtained for $\phi(d)$, once the substitution for $\sigma$ has been made.

The reliability of this approximation may be tested by calculating $r_{S'\phi}$ from the available data. The 41 tests gave r = .867, and the 22 tests r = .817 .

As already indicated, the process of approximation may be carried a stage further. In the expression for $\phi(d)$, put $d = \frac{1}{2} + \delta$ . Then

$$\phi(d) = \frac{2d-1}{\sqrt{d(1-d)}} = \frac{4\delta}{\sqrt{1-4\delta^2}} = 4\delta \left(1 + 2\delta^2 + \ldots\ldots\right).$$

Now $|\delta| < \frac{1}{2}$ ; as a first approximation we may therefore take

$$\phi(d) = 4\delta = 4d - 2,$$

which is a linear function of d.

The error introduced by this latter approximation varies with the size of $\delta$ , that is with the value of d. For d equal to 0.3 or 0.7, i.e., $\delta = \pm 0.2$, the error is about 10%, and it decreases as $|\delta|$ decreases.

Within these limits, $\phi$ is practically a linear function of d. The great majority of the tests under consideration had difficulty levels within this range. This is probably the reason for the apparent anomaly of S being correlated in linear fashion with such varying functions of d as $d^3$, $\phi(d)$, and d itself.

An interesting commentary on the above approximation is provided by the regression equation predicting S' from d, as calculated from the data of the 41 tests. When this equation is put into the form where S' and d are measured in their original units it becomes $S' = 4.10\,d - 2.07$ .

This long and rather involved discussion therefore leads to the conclusion that the difficulty level d is a first approximation to $\phi(d)$, which in turn is a first approximation to S', the skewness of the answer-pattern-differential. The correlation found to exist between S, the skewness of the score-scatter, and d, the difficulty level, is an effect of the correlation existing between S and S'.

## 10. Conclusion.

The wellknown rule-of-thumb method of skewing score-scatters by the variation of difficulty level is an approximation to the basic principle, that the skewing of score-scatters is controlled by the skewness of the answer-pattern-differential. Since it is an approximation, this method cannot be expected to yield such accurate results as that using the answer-pattern-differential; this is demonstrated by the diminution in the

of their array; the other the variation of the means of the arrays about the general mean. The variation of the means about the general mean is partly due to the slope of the regression line, and the amount of this variation is calculable; the other part is due to deviations from linearity. For each variation the standard deviation is calculable, and the test of significance of the deviations from linearity becomes the test of whether the standard deviation of the deviations from linearity is, or is not, significantly greater than the standard deviation of the observations within the array about their mean.

In the correlation table on page 99, number the arrays from left to right, omitting the array in which there are no entries. For ease of calculation, an assumed mean may be taken at S = -0.35, and the class intervals are taken as unit. Then for each array of S, the excess is calculated, and this is also done for the total. The results of these calculations are shown in the table overleaf.

The sum of the mean squared excess less the mean square of the total excess gives $\sum n_p (\bar{S}_p - \bar{S})^2$ , which here equals 300.8 .

The total variation of S has already been found in the calculations of the correlation; it is 532.9 . Thus the variation $\sum\sum( S - \bar{S}_p )^2$ equals 532.9 - 300.8 = 232.1 .

correlations shown in the table below.

Table 27. Correlation of S with S', $\phi$, and d.

| | 41 tests | 22 tests |
|---|---|---|
| $r_{SS'}$ | .628 | .836 |
| $r_{S\phi}$ | .576 | .816 |
| $r_{Sd}$ | .561 | .790 |

For this reason a better prediction of S will always be obtained from S' than from d alone. From a study of the whole answer-pattern one gains a better idea of the probable skewness of the score-scatter than would be obtained from a knowledge of the test's difficulty alone.

Appendix. Fisher'S Method of testing Linearity of Regression.

Since this method is of fairly recent origin, an example is worked out in full below, the regression tested being that of S on d for the 41 tests. The correlation table is on page 99.

For each value of d there is an array of values of S. Let us designate any array by the suffix p; the number of observations in the array may be denoted by $n_p$, and the mean of the array by $\bar{S}_p$; the mean of all the values of S is denoted by $\bar{S}$. Then it may be shown that

$$\sum( S - \bar{S} )^2 = \sum\sum( S - \bar{S}_p)^2 + \sum\{n_p ( \bar{S}_p - \bar{S} )^2\}.$$

This is an algebraic identity, expressing the fact that the total variation of S may be split up into two parts, one representing the variation of the observations about the mean

of their array; the other the variation of the means of the arrays about the general mean. The variation of the means about the general mean is partly due to the slope of the regression line, and the amount of this variation is calculable; the other part is due to deviations from linearity. For each variation the standard deviation is calculable, and the test of significance of the deviations from linearity becomes the test of whether the standard deviation of the deviations from linearity is, or is not, significantly greater than the standard deviation of the observations within the array about their mean.

In the correlation table on page 99, number the arrays from left to right, omitting the array in which there are no entries. For ease of calculation, an assumed mean may be taken at S = -0.35, and the class intervals are taken as unit. Then for each array of S, the excess is calculated, and this is also done for the total. The results of these calculations are shown in the table overleaf.

The sum of the mean squared excess less the mean square of the total excess gives $\sum n_p (\bar{S}_p - \bar{S})^2$ , which here equals 300.8 .

The total variation of S has already been found in the calculations of the correlation; it is 532.9 . Thus the variation $\sum\sum( S - \bar{S}_p )^2$ equals 532.9 - 300.8 = 232.1 .

Table 28. Calculation of $\sum n_p ( \bar{S}_p - \bar{S} )^2$.

Arrays of S for given values of d.

| S | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -8 | | | | 1 | | | | | | | | | | 1 |
| -7 | | | | | | | | | | | | | | 0 |
| -6 | | | 1 | 1 | | | | | | | | | | 2 |
| -5 | | | | | | | | | | | | | | 0 |
| -4 | | 2 | 1 | | | | 1 | | | | | | | 4 |
| -3 | | | 1 | | | | | | 1 | | | | | 2 |
| -2 | | | 1 | | | | | | 1 | | | | | 2 |
| -1 | | | | 1 | | | 1 | | | | | | | 2 |
| 0 | 1 | | | 1 | 1 | | | | | | | | | 3 |
| 1 | | | 1 | 2 | 2 | 1 | | 1 | | | | | 1 | 8 |
| 2 | | | | | 1 | 1 | | 1 | | 1 | | | | 4 |
| 3 | | | | | | 1 | | 1 | | 1 | | | | 3 |
| 4 | | | | | 1 | 1 | 2 | | | | | | | 4 |
| 5 | | | | | | | | | 1 | 1 | 1 | | | 3 |
| 6 | | | | | | | | 1 | | | | | | 1 |
| 7 | | | | | | | | 1 | | | | 1 | | 2 |
| $n_p$ | 1 | 2 | 5 | 6 | 5 | 4 | 4 | 5 | 3 | 3 | 1 | 1 | 1 | 41 |
| Excess | 0 | -8 | -14 | -13 | +8 | +10 | +3 | +19 | 0 | +10 | +5 | +7 | +1 | +28 |
| $\frac{(\text{Excess})^2}{n_p}$ | 32 | 392 | 282 | 128 | 25 | | 22 | 722 | 0 | 333 | 25 | 49 | 1 | 19.1 |

Total = 319.9 - 19.1 = 300.8

For each of these terms there must also be found the number of degrees of freedom. The number of observations is 41, therefore the number of degrees of freedom for the total variation is 40. The number of arrays is 13, therefore the number of degrees of freedom for the variation between arrays is 12. By subtraction the number of degrees of freedom within the arrays must be 28 .

Of the variation between the arrays part is due to the slope of the regression line. This may easily be shown to be equal to $r^2 \sum ( S - \bar{S} )^2$, which in the present case is 167.8 . The number

of degrees of freedom represented here is 1.

The variation between the arrays due to deviations from linearity is therefore equal to 300.8b- 167.8 = 133.0, and the corresponding number of degrees of freedom is 12 - 1 = 11 . We have therefore to decide whether a variation of 133.0 obtained from 11 degrees of freedom is significantly greater than a variation of 232.1 obtained from 28 degrees of freedom.

This is the same problem as determining whether an estimate of standard deviation derived from $n_1$ degrees of freedom is significantly greater than a second estimate obtained from $n_2$ degrees of freedom. The method is to evaluate z equal to the difference of the natural logarithms of the two standard deviations, i.e., $z = \log_e \frac{\sigma_1}{\sigma_2}$ ; then the probability of exceeding this value of z by chance is tabulated in Fisher's Table VI for given values of $n_1$ and $n_2$ . As before, a probability of .05 is taken as the dividing line.

In the present example the calculation may be completed thus

| Variance of S | | Degrees of freedom | Mean square | $\frac{1}{2}\log_e$ |
|---|---|---|---|---|
| Total | 532.9 | 40 | | |
| Between arrays | 300.8 | 12 | | |
| ∴ Within arrays | 232.1 | 28 | 8.29 | 1.06 |
| Due to linear regression | 167.8 | 1 | | |
| ∴ Due to deviations | 133.0 | 11 | 12.09 | 1.24 |
| | | | | z = 0.18 |

For $n_1$ = 11, $n_2$ = 28, the 5% point is 0.39; that is, the probability of exceeding z = 0.39 by chance is 1 in 20. The regression of S on d, with z = 0.18 is definitely linear.

## Chapter 8. The Measurement of the Control of Score-scatter by Answer-pattern in single tests.

### 1. The nature of the problem.

In chapter 2 it was shown that the answer-pattern-differential and the score-scatter of a test were identical in the particular case referred to as "unig", where each candidate's score was compiled of answers to the easiest possible questions. The results submitted in chapters 5, 6, and 7 show that in the more general case, where randomness of answering is present, there is still some measure of relation or correspondence between the two distributions, though there is no longer identity. The measure of correspondence in each case was the correlation coefficient of corresponding statistics of the distribution. The calculation of these correlations implies the existence of several, and if possible of many, tests given to the same or similar populations.

The problem to be studied in this chapter is the measurement of the degree of correspondence of the two distributions in a single test. Given the results of a single test, is it possible to estimate the degree of control the answer-pattern has exerted on the score-scatter ?

First it may be pointed out that the answer-pattern-differential and score-scatter are so removed from identity that no use can be made of Pearson's test of Goodness of Fit. In practically every case this test when applied to the answer-pattern-differential

and score-scatter of a test, would indicate these to be entirely different distributions. In any case, Pearson's test can only be used to decide whether or not two distributions may be regarded as identical save for the effects of sampling, and cannot be used to measure the degree of control or goodness of fit of the two, as the name might suggest.

It is necessary, then, to devise some other method of measuring the relationship that undoubtedly exists even in the case of a single test. The problem is so beset with difficulties that, although four coefficients have been devised, evaluated for all the tests, and otherwise used, none appears satisfactory. Only two will be mentioned here.

2. The coefficient of hig -"h"

This coefficient was devised and used by the author in studies for the degree of Bachelor of Education.

When there is exact correspondence between answer-pattern-differential and score-scatter, the equations (B) of chapter 2 hold. That is,

$n_x - n_{x+1} = N_x$ holds for x = 0, 1, 2,.....m.

The extent of deviations from this state may then be measured by the expression

$$\sum_{x=0}^{m} ( n_x - n_{x+1} - N_x )^2$$

which obviously vanishes when equations (B) hold.

Conversely, since the expression is the sum of squares, when it vanishes each term must vanish; i.e.

$n_x - n_{x+1} = N_x$ is true for x = 0, 1, 2, .....m,

that is there is exact correspondence between answer-pattern-differential and score-scatter. The equations also imply that the test is unig, i.e. that each candidate's score must be made up of answers to the easiest questions. For, the last equation is $n_m = N_m$, i.e. those, and only those, making the maximum score have answered the last question. Since $n_{m-1} - n_m = N_{m-1}$ or $n_{m-1} = n_m + N_{m-1}$, the penultimate question must have been answered by $n_m + N_{m-1}$ candidates. Now these include the $N_m$ candidates who made perfect scores, and also the $N_{m-1}$ who scored m-1, since these latter could not have answered question m as an alternative. Therefore the number of times question m-1 is answered is accounted for entirely by those candidates scoring m-1 and upwards. A similar argument extends to scores m-2 and so on. The equation $\sum ( n_x - n_{x+1} - N_x )^2 = 0$ therefore implies the unig type of answering.

The expression $\sum ( n_x - n_{x+1} - N_x )^2$ has a minimum value of zero when the scores are unig. It is at a maximum when the scores are made in random or hig fashion. The probability of this, as was proved in chapter 2, is greatest when the answer-pattern is flat, i.e. when $n_1 = n_2 = n_3 = ..... = n_m = n$ say.

Then the sum becomes

$$\sum ( n_x - n_{x+1} - N_x )^2 = \sum_{o}^{m} N_x^2 + n^2 + (n_o - n)^2 - 2nN_m - 2(n_o - n)N_o$$

To obtain a coefficient which will make it possible to compare tests with differing numbers of candidates and differing

numbers of items, let us define the coefficient of hig as

$$h = \frac{\sum(n_x - n_{x+1} - N_x)^2}{\sum N_x^2 + n^2 + (n_o - n)^2 - 2nN_m - 2(n_o - n)N_o}$$

where n is the mean of $n_1$, $n_2$, .....$n_m$.

This coefficient is independent of the number of candidates sitting the test; for if the candidates be augmented by a similar population the values of n and N will be increased in the same ratio, and the expression for h being homogeneous in n and N remains the same. The effect of varying the number of items is rather complicated and will not be investigated here. In actual results from tests it has been found that the coefficient calculated directly from the data of a 100 item test does not differ much from the coefficient obtained from the same data after grouping the scores and answer-pattern-differentials into 10 groups of 10, i.e., replacing the 100 item test by a 10 item test.

The method of calculation may be illustrated from the data of Test 4 of the 41 tests, already used as an example in previous chapters.

Table 29. Calculation of coefficient of hig h .

| x | $n_x$ | $n_x - n_{x+1}$ | $N_x$ | $N_x^2$ | $\lvert n_x - n_{x+1} - N_x \rvert$ | $\lvert\rvert^2$ |
|---|---|---|---|---|---|---|
| 0 | 32 | 3 | 1 | 1 | 2 | 4 |
| 1 | 29 | 2 | 1 | 1 | 1 | 1 |
| 2 | 27 | 4 | 1 | 1 | 3 | 9 |
| 3 | 23 | 5 | 10 | 100 | 5 | 25 |
| 4 | 18 | 6 | 5 | 25 | 1 | 1 |
| 5 | 12 | 3 | 4 | 16 | 1 | 1 |
| 6 | 9 | 1 | 4 | 16 | 3 | 9 |
| 7 | 8 | 1 | 2 | 4 | 1 | 1 |
| 8 | 7 | 1 | 2 | 4 | 1 | 1 |
| 9 | 6 | 0 | 2 | 4 | 2 | 4 |
| 10 | 6 | 6 | 0 | 0 | 6 | 36 |
| | 145 | | | 172 | | 92 |

$n = 145/10 = 14.5$ $\quad \Sigma N^2 = 172$

$n^2 = 210$

$n_o - n = 17.5$ $\quad (n_o - n)^2 = 306$

$688$

$N_o = 1$ $\quad 2(n_o - n)N_o = 35$

$653$

$N_m = 0$

$h = 92/653 = 0.14$

---

The values of h in the 41 tests ranged from .06 to .41, the median being .24 . In the complete tests h ranged from .06 to .40, the values for some of these tests were

| | |
|---|---|
| M.H.T. 8 | .25 |
| M.H.T. 9 | .11 |
| M.H.T. 11 | .10 |
| M.H.T. 12v | .16 |
| M.H.T. 12p | .20 |
| Thesis A | .40 |
| Thesis B | .09 |
| Thesis C | .07 |

3. Criticism of h.

There are two weak points in the definition of h. First, in the process of obtaining the denominator representing the incidence of maximum hig, the answer-pattern of the test was altered to a flat answer-pattern of the same difficulty level. The score-scatter was left unchanged. Now the evidence of chapter 4 indicates that the answer-pattern of a test is just as permanent a feature as the score-scatter; in altering the answer-pattern we have changed the whole test.

At bottom, this weakness seems to depend on a confusion of ideas between hig as poorness of fit of answer-pattern-differential and score-scatter, and hig as randomness of answering due to lack of agreement between the individual examinee's order of difficulty of items and the average order of difficulty. These two ideas are of course linked by the unig case where there is perfect fit and perfect agreement.

The second weak point is as fundamental. It has already been noted that there is a natural tendency of score-scatters to normality. An answer-pattern-differential which is already of this shape will apparently have a better chance of showing a good fit to any score-scatter obtained than would an answer-pattern-differential of a completely different shape, say the answer-pattern-differential of a flat test. Thus the low value of h obtained with tests M.H.T. 9, M.H.T. 11, Thesis B and C may be attributed to the shapes of their answer-patterns, givin

answer-pattern-differentials already akin to normal distributions. On the other hand the high value of h obtained from test A may be attributed to the shape of A's answer-pattern-differential, which was of the type shown in figure 22.

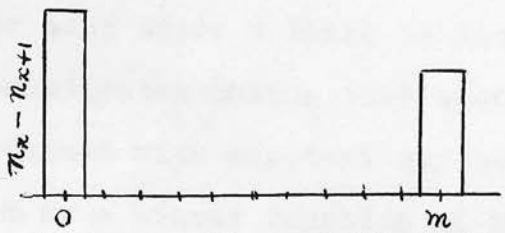



Figure 22. Answer-pattern-differential of a flat test.

Although h is large in the case of test A, it may be that the answer-pattern of the test actually exerted more control in the production of the score-scatter than did the answer-patterns of tests B and C, but the effect is masked by tendency of score-scatters to normality.

4. Definition of $\alpha$.

In the formation of the score-scatter of a test there seem then to be two factors, the answer-pattern-differential of the test, and what might be called the "natural" distribution of the scores. This is the distribution that would be obtained if the influence of the answer-pattern could be removed. From what we have learned we should postulate it as a normal distribution, which implies that it is unskewed; and its mean is of course fixed as equal to the mean of the answer-pattern-differential, or what is the same thing, the mean of the score-scatter formed. Though the mean and skewness of the distribution are so fixed,

there is no method of estimating its standard deviation, even when the tests are mental tests and the standard deviation of intelligence quotients is taken as 13 points.

Let us suppose meantime that this distribution is known; that is for each score x there is known the value of $w_x$, the number of candidates making that score. Then the actual score-scatter obtained with anyvtest may be regarded in its most simple form as a linear function of the answer-pattern-differential and the natural score-scatter; that is

$$N_x = \alpha (n_x - n_{x+1}) + \beta w_x$$

where $\alpha, \beta$ are parameters. The ratio $\alpha/\beta$ is a measure of the relative influence of answer-pattern-differential and natural score-scatter in the formation of N.

In a ten item test there are eleven such equations, the values of $\alpha$ and $\beta$ varying with the equation. These may be regarded as eleven equations for $\alpha$ and $\beta$, and from them have to be derived single values of $\alpha$ and $\beta$ which best represent the position for the whole test. The number of equations is too small to enable us to set up a regression equation determining the best values of $\alpha$ and $\beta$; in any case, the variables $N_x$, $n_x - n_{x+1}$ and $w_x$ are not normally distributed.

The best values of $\alpha$ and $\beta$ may be determined by the method of least squares. Each equation is multiplied by the coefficient of $\alpha$, and the resulting equations are added to give one equation in $\alpha$ and $\beta$. Similarly, by multiplying each

equation by the coefficient of $\beta$ and adding, a second equation in $\alpha$ and $\beta$ is obtained. The solution of these two simultaneous equations gives the values of $\alpha$ and $\beta$ which fit best all the original equations. If $n_x - n_{x+1}$ is denoted for brevity by $v_x$ the equations may be written

$$\alpha = \frac{\sum Nv \sum w^2 - \sum Nw \sum vw}{\sum v^2 \sum w^2 - (\sum vw)^2}$$

$$\beta = \frac{\sum Nw \sum v^2 - \sum Nv \sum vw}{\sum v^2 \sum w^2 - (\sum vw)^2}$$

In the application of this formula to the 41 tests the first difficulty is the construction of the w distributions appropriate to each test. As noted before, the mean and skewness of these distributions is fixed, but some value for the standard deviation must be assumed before the distribution can be calculated. Since no theoretical basis has so far been found on which the standard deviation could be estimated, some rule of thumb method must be used meantime. The following method was that finally chosen for the 41 tests. It was found that the average standard deviation of the scores in these tests was 2.12 . This also approximates closely to the standard deviation of scores that would most probably have been obtained from a test of that type, the items of which progressed uniformly in difficulty from very easy to very hard. In such a test it is easy to show that the standard deviation of the answer-pattern-differential is 3.16 , and the corresponding standard deviation of scores obtained from the regression equation for the 41 tests is 2.10. A value near 2.10

had therefore both a theoretical and practical significance for the 41 tests. To ease the calculation of w somewhat, the standard deviation finally chosen was 2.00 .

Thereafter the method of calculation followed closely that demonstrated in the appendix to chapter 2. As a sample, there is shown below the table worked out for test 4. The scores were grouped in twos to decrease random errors due to the smallness of the frequencies in some of the classes.

The mean score, obtained from the answer-pattern-differential before grouping, was 4.5 .

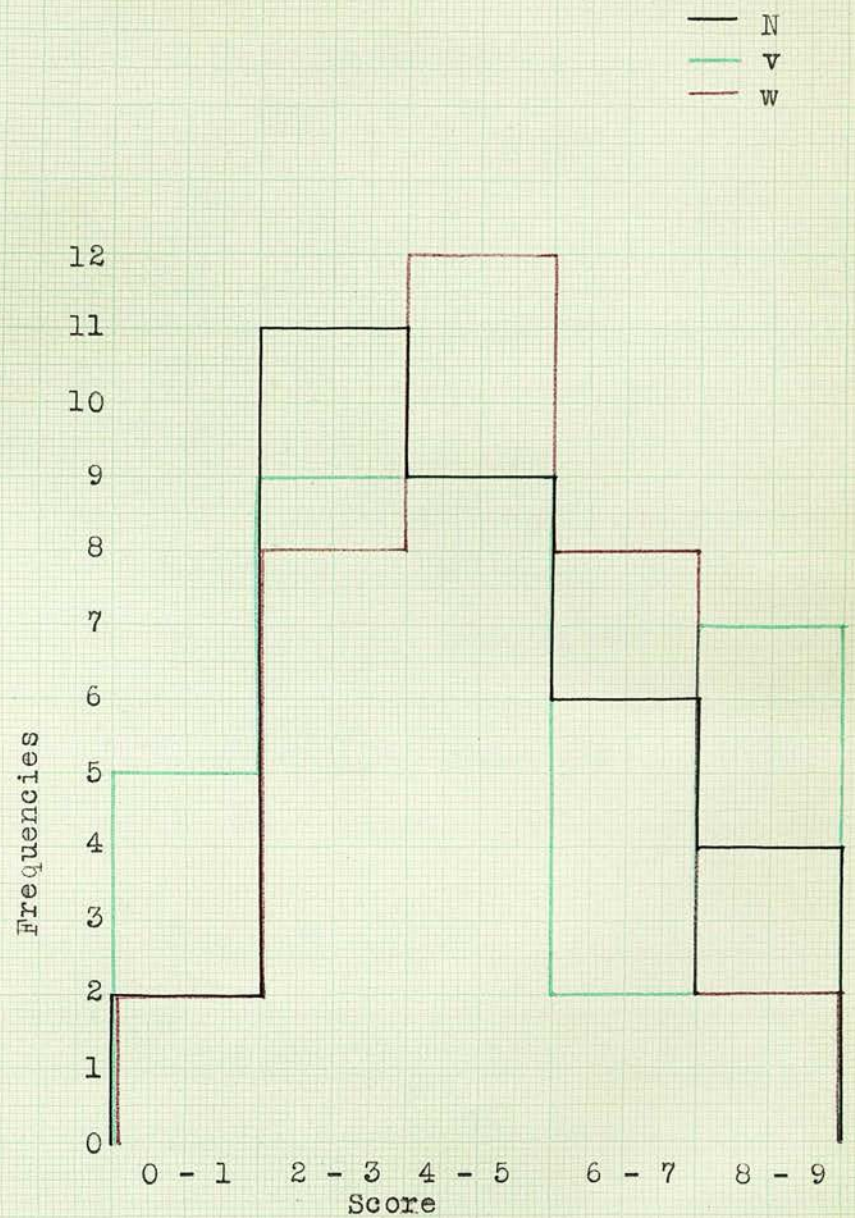
Table 30. Calculation of w distribution for test 4.

| Score | x | $\frac{x-a}{\sigma}$ | I | Difference | 32 × Difference |
|---|---|---|---|---|---|
| | | $[-\infty]$ | -.500 | | |
| | | | | .067 | 2 |
| 0-1 | 1.5 | -1.5 | -.433 | | |
| | | | | .242 | 8 |
| 2-3 | 3.5 | -0.5 | -.191 | | |
| | | | | .382 | 12 |
| 4-5 | 5.5 | +0.5 | +.191 | | |
| | | | | .242 | 8 |
| 6-7 | 7.5 | +1.5 | +.433 | | |
| | | | | .067 | 2 |
| 8-9 | | $[+\infty]$ | +.500 | | |
| | | | | | 32 |

For test 4 the normal distribution was then taken as above, 2,8,12,8,2 for the appropriate scores.

Now the use of an answer-pattern-differential 5,9,9,2,7 altered the above normal score-scatter to that actually observed i.e., 2,11,9,6,4. These distributions are shown in histogram form overleaf ( page 127 ).

Fig. 21. Distributions N, v, and w for Test 4.



The calculation of $\alpha$ and $\beta$ then proceeds as follows.

Table 31. Calculation of $\alpha$ and $\beta$ for test 4.

| Score | N | v | w | $v^2$ | $w^2$ | Nv | Nw | vw |
|---|---|---|---|---|---|---|---|---|
| 0-1 | 2 | 5 | 2 | 25 | 4 | 10 | 4 | 10 |
| 2-3 | 11 | 9 | 8 | 81 | 64 | 99 | 88 | 72 |
| 4-5 | 9 | 9 | 12 | 81 | 144 | 81 | 108 | 108 |
| 6-7 | 6 | 2 | 8 | 4 | 64 | 12 | 48 | 16 |
| 8-9 | 4 | 7 | 2 | 49 | 4 | 28 | 8 | 14 |
| Totals | 32 | 32 | 32 | 240 | 280 | 230 | 256 | 220 |

$\alpha = 0.43$, $\beta = 0.58$

In test 4 then, the relative influences of answer-pattern-differential and of tendency to normality are in the ratio 43/58.

As it is found that $\alpha + \beta$ approximates closely to unity in every case, the value of $\alpha$ alone may be given as a sufficient indication. It was found in the 41 tests that the value of $\alpha$ ranged from -0.22 (test 24) to +0.50 (test 39) with a median value +0.14 .

The weaknesses of this method of determining the relative strengths of answer-pattern-differential and normal tendency are obvious. It demands a precise knowledge of the "natural" distribution which we do not have at the present stage. In its calculation there arise difficulties such as the calculation of w for tests with high or low means, causing lack of headroom or the opposite. The theoretical basis of the natural distribution is flimsy. In referring to a score-scatter uninfluenced by any answer-pattern-differential we are creating a mathematical fiction.

## 5. Conclusion.

These difficulties bring us to consider the necessity of constructing such coefficients as h and $\alpha$. Their usefulness is based on the following train of reasoning. There has been proved to be a certain measure of agreement between the answer-pattern-differential of a test and its score-scatter. The examiner who can select his test items from a battery of items of known difficulty can therefore predict within certain limits the nature of the score-scatter that will be obtained when the test is given to a known population. It may be that tests vary in the extent of the agreement between their answer-pattern-differential and score-scatter ( though that is by no means definitely proved yet), some tests showing closer agreement, and others a greater tendency for the score-scatter to vary. Then the examiner would tend to choose that type of test for which the agreement was good so that his attempt to procure a given score-scatter would be more likely to be successful. In defining h, $\alpha$ and other coefficients not mentioned here, we are seeking some way of labelling tests so that, having divided the sheep from the goats, we may analyse the factors causing some to be sheep and the others goats. It may be, as stated above, that there is no such distinction; that the fluctuations in agreement are due merely to errors of sampling, and the small numbers of tests that have perforce had to be used render this possibility quite feasible. More light is shed on the problem by the results of the next chapter.

## Chapter 9. The Relation of Steepness of Tests to the Control of Score-scatter by Answer-pattern.

### 1. The definition of steepness.

The problem of what type of test gives the best agreement of answer-pattern-differential and score-scatter may be tackled in a different way from that adopted in the last chapter. It was shown in chapter 2 that the probability of unig was greatest when the test was of the steep type, i.e. one in which $n_1$ is much greater than $n_2$, which in turn is much greater than $n_3$, and so on. Since unig implies perfect fit of score-scatter and answer-pattern-differential it would seem a priori that the steeper tests should show closer agreement of score-scatter and answer-pattern-differential than would be obtained with flatter tests.

The definition of "steep" is so far no more than the bare statement that in such a test $n_1$ is much greater than $n_2$, $n_2$ than $n_3$, and so on. What this implies may conveniently be considered in two steps.

(1) It is obvious that steepness depends to a great extent on the number of items in the test. A test with few items has, a priori, a much greater chance of being made steep than a test with many items, since in the latter case it is impossible to make great differences of difficulty between adjacent items. Consider, for example, the following three item test.

Simplify (1) $2 + 3 - 6$

(2) $\frac{4}{3}\left\{\frac{5}{8} + \frac{1}{2}\right\}$

(3) $\frac{ab}{a+b}\left\{\frac{a}{b} + \frac{b}{a}\right\}$

This is a steep test. It is almost certain that the answers to such a test would be of unig type; any candidates answering correctly question 3 would answer correctly questions 1 and 2, and so on. The introduction of additional items of intermediate difficulty would obviously lessen the degree of certainty, and if the number of items were increased to ten, even though these made use of the full range of difficulty between item 1 and item 3 of the above test, there would obviously be little hope of the answers being completely unig.

This state of matters is reflected in the probability of hig, as defined on page 32. For a test of three items, with an answer-pattern $n_0 = 100$, $n_1 = 80$, $n_2 = 50$, $n_3 = 20$, the probability of unig is 0.6 . For a very similar test with the same general shape of answer-pattern but having four items, $n_1 = 80$, $n_2 = 60$, $n_3 = 40$, $n_4 = 20$, the probability of unig is only 0.2 . For a ten-item test of any shape of answer-pattern, the probability of unig is so low as to be negligible.

(2) Although the number of items to be used may be so large that the probability of unig is negligible, it may be that tests with steeper answer-patterns,(i.e. with difficulty differences as large as possible in view of the number of items present )

show closer agreement between the answer-pattern-differential and the score-scatter. These tests will be referred to as "steeper" rather than as "steep". For a study of this it will be necessary to define more exactly what is meant by steepness in the case of tests with equal numbers of items.

Consider the answer-patterns illustrated in the diagram.

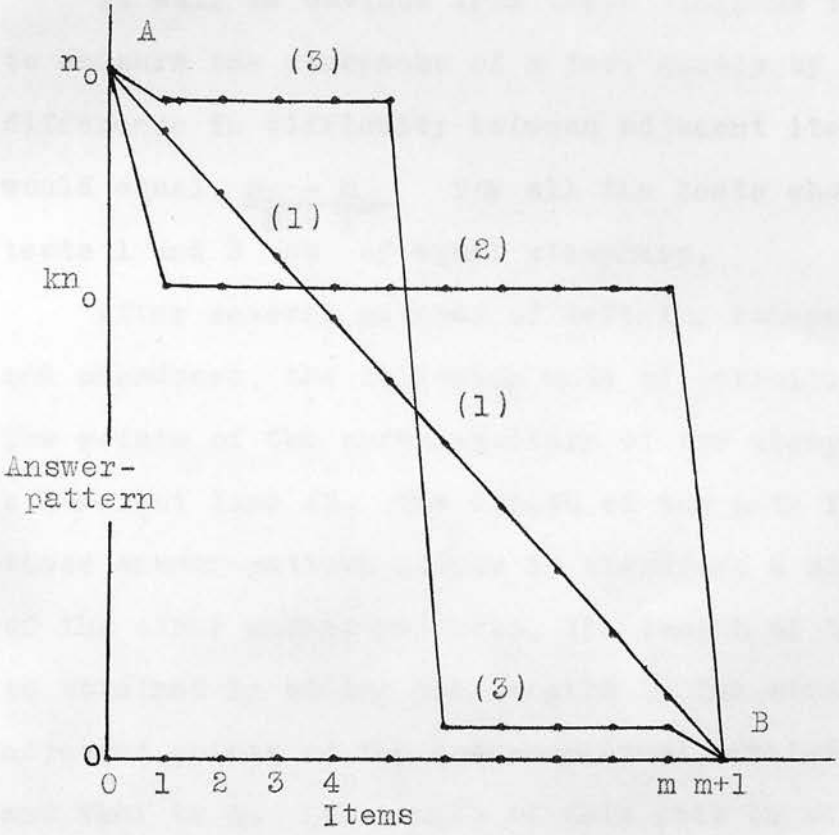



Figure 24. Answer-patterns of various types of test.

(1) is the answer-pattern of what might be called the steepest possible test for the given number of items. The points lie on a straight line joining $x=0$, $y=n_o$ to $x=m+1$, $y=0$. Strictly the curve making the probability of unig a maximum is not a straight line, but the whole calculation is of such mathematical

difficulty, and the result so far as obtained so nearly a straight line, that this has been adopted.

(2) is the answer-pattern of a flat test, each item of which has been answered by $km_o$ of the $n_o$ candidates. $(k<1)$

(3) represents an answer-pattern partly steep and partly flat.

It will be obvious from these diagrams that it is impossible to measure the steepness of a test merely by averaging the difference in difficulty between adjacent items. Such an average would equal $\frac{n_1 - n_m}{m - 1}$ for all the tests shown, and would rank tests 1 and 3 as of equal steepness.

After several methods of defining steepness had been tried and abandoned, the following mode of definition was devised. The points of the answer-pattern of the steepest test (1) lie on a straight line AB. The length of the path from A to B through these answer-pattern points is therefore a minimum. In the case of the other answer-patterns, the length of the path from A to B is obtained by adding the lengths of the straight lines joining adjacent points of the answer-pattern, including the line from A and that to B. The length of this path is an index of the flatness of the test.

In test (1), the length of the path, l, equals $\sqrt{n_o^2 + (m+1)^2}$ in test (2) it equals $\sqrt{(1-k)^2 n_o^2 + 1} + (m-1) + \sqrt{k^2 n_o^2 + 1}$ which is approximately equal to $(1-k)n_o + m - 1 + km_o = n_o + m - 1$ independent of the value of k. This may be taken as the

maximum length of the path, but for our purpose it is even better to take $n_0 + m + 1$ as the maximum length. The advantage of this choice is that it makes easier the construction of a coefficient which will be independent of the number of candidates sitting the test. For the same reason it is necessary to fix the scale of the diagram by taking as a representative number of candidates $m + 1$. This makes the triangle AOB isosceles and the minimum length of the path becomes $(m+1)\sqrt{2}$, and the maximum length $2(m+1)$.

Suppose that the original answer-pattern of any test has the values $n_0, n_1, n_2, \ldots\ldots n_m$, and denote the corresponding answer-pattern-differentials $n_x - n_{x+1}$ by $\Delta_x$. Then, after adjustment of the diagram scale to make the triangle isosceles, the differences will be $\frac{m+1}{n_0}\Delta_x$, and the length of the path from A to B will be $\sum_{x=0}^{m}\sqrt{1+\left(\frac{m+1}{n_0}\Delta_x\right)^2}$.

The coefficient of steepness $c$ may now be defined as the ratio*

$$c = \frac{\text{maximum length - actual length}}{\text{maximum length - minimum length}}$$

$$= \frac{2(m+1) - \sum\sqrt{1+\left(\frac{m+1}{n_0}\Delta_x\right)^2}}{2(m+1) - \sqrt{2}(m+1)}$$

$$= \frac{1}{.293}\left\{1 - \frac{1}{2(m+1)}\sum\sqrt{1+\left(\frac{m+1}{n_0}\Delta_x\right)^2}\right\}$$

The calculation of this coefficient for any of the 41 tests is a very simple matter. It is shown for test 11 overleaf.

*So defined to make $c = 0$ for a flat test, and $c = 1$ for the steepest test.

Table 32. Calculation of c for test 11.

| Answer-pattern | A.P.D. $\Delta$ | $\frac{11}{32}\Delta$ | $\sqrt{1+\left(\frac{11}{32}\Delta\right)^2}$ |
|---|---|---|---|
| 32 | 5 | 1.72 | 1.99 |
| 27 | 1 | .34 | 1.08 |
| 26 | 0 | 0 | 1.00 |
| 26 | 2 | .69 | 1.21 |
| 24 | 0 | 0 | 1.00 |
| 24 | 1 | .34 | 1.08 |
| 23 | 0 | 0 | 1.00 |
| 23 | 0 | 0 | 1.00 |
| 23 | 1 | .34 | 1.08 |
| 22 | 9 | 3.10 | 3.26 |
| 13 | 13 | 4.47 | 4.58 |
| | | | 18.28 |

$$c = \frac{1}{.293}\left(1 - \frac{18.28}{22}\right) = 0.6$$

This coefficient was calculated for each of the 41 tests, and the values obtained ranged from 0.6 (test 11) to 0.9 ( test 20 ).

3. The influence of steepness on the fit of answer-pattern-differential and score-scatter.

By the use of this measure of steepness, the 41 tests may be divided into two groups of steeper and flatter tests, the dividing line being fixed so as to place approximately equal numbers of tests in each group. Each of the groups then yields data from which can be calculated the correlation of the standard deviations of the answer-pattern-differential and the score-scatter (as was done in chapter 6 for the 41 tests), and the correlation of the coefficients of skewness of these distributions ( as was done in chapter 7). When these were

calculated the results were as follows.

The correlation of the standard deviations of answer-pattern-differential and of score-scatter for the 20 steeper tests was r = .776; for the 21 flatter tests it was r = .830. This difference is not significant, the difference of the corresponding values of z being .15 and its standard deviation being .34 . In the matter of standard deviations the flatter tests thus show a fit which is better , but not significantly so, than do the steeper tests.

The correlation of the skewness of the answer-pattern-differential with the skewness of the score-scatter for the 20 steeper tests was r = .795; for the 21 flatter tests it was r = .590. This difference, though greater than that found with the standard deviations, is still not significant: the difference in the values of z is .41, and its standard deviation is .34 . Here the better fit is obtained from the steeper tests.

From the scanty data at our disposal it would seem then that the control of score-scatter by answer-pattern-differential is no greater in the case of the steeper group than in the case of the flatter group. This conclusion is strengthened by the result obtained in a second method of examining the position in these 41 tests.

The coefficient $\alpha$ defined in chapter 8, measured the relative influence of the answer-pattern-differential and a hypothetical normal distribution on the score-scatter obtained. By correlating this coefficient with the coefficient c we may

ascertain whether the steeper tests show a relatively greater influence of answer-pattern-differential. This correlation was calculated from the data of the 41 tests, and was found to be $r = +.218$, which though positive is not significantly different from zero, being derived from 41 pairs of values.

It would appear then that for tests of ten items or more, the control of score-scatter by answer-pattern-differential is independent of the steepness of the test. In a way, this is rather a welcome conclusion, since it simplifies matters somewhat As will be shown in the next chapter, it is necessary to use flat answer-patterns to produce certain types of score-scatter, If the certainty of control decreased with the flatness of the answer-patterns the position would be much more confused than it is. As far as these results go, they show that a flat answer-pattern exerts about as much control over the score-scatter as does any other type.

On the other hand the results show that to be really steep in the sense of producing answers approaching unig type, a test must have very few items, say three or four, and these spaced out in the best way. This is a type of test few examiners would care to use; it would very easily give rise to difficulties through misunderstandings, prior knowledge, and guesswork, all of which may be smoothed out in a test with many items.

We are thus in the position of having "gained upon the roundabouts what we lost upon the swings". We have lost hope of constructing useful unig tests, but have gained the knowledge that

flat tests show as great a correspondence between answer-pattern and score-scatter as do the steeper tests.

4. An advantage in the use of steeper tests.

There is one advantage of the steeper test which is worth mentioning. It sometimes happens that a test designed for a group of given average ability is employed to test a group of slightly different ability. It is easy to show by an example that in such a case the steeper test has its characteristics changed less by the altered character of the testees than has a flat test of the same average difficulty.

Consider two tests which when applied to a certain group of candidates yield the following answer-patterns.

Test 1

| Item | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 100 | 91 | 82 | 73 | 64 | 55 | 46 | 37 | 28 | 19 | 10 |

This represents the steepest possible test of ten items. Tha answer-pattern-differential has an average score 5, a standard deviation 3.16 and skewness 0.00.

Test 2

| Item | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 100 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |

This represents a flat test, with an average score 5, a standard deviation of answer-pattern-differential 5.00 and skewness again 0.00 .

Suppose now that these two tests are given to a group of 100 candidates, whose mean ability is less than that of the original group by an amount sufficient to depress the percentage of correct answers to the items of test 2 from 50 to 40. This corresponds, in the case of an intelligence test, to an age difference of four months at age eleven. Using the technique described in chapter 4 we can now calculate the changes in the answer-patterns of both tests.

The answer-pattern of test 1 becomes

100, 86, 75, 65, 55, 45, 37, 28, 20, 13, 5,

yielding a mean score 4.29, and an answer-pattern-differential with standard deviation 3.12 and skewness +0.30.

The answer-pattern of test 2 becomes

100, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40.,

yielding a mean score 4.00, and an answer-pattern-differential with standard deviation 4.90 and skewness +0.41 .

It will readily be seen that in each respect the steeper test has suffered less change than the flat one. This superiority of the steeper test in permanence of answer-pattern will tend to be reproduced in the score-scatters, since these are in part controlled by the answer-patterns.

Another feature of the steeper tests is that their reliability as measured by the split-halves method is likely to be greater This question will be discussed in a later chapter.

## Chapter 10. The Construction of Answer-patterns.

### 1. The theoretical basis.

From the preceding theory it is apparent that the examiner who wishes to produce a score-scatter of a given type must construct an answer-pattern appropriate to his purpose. This assumes that he has access to a battery of items of known difficulty for the population considered: such a collection has been made, for instance, by Professor Thorndike and is described in his "Measurement of Intelligence", and a similar collection might be compiled from the data of the Moray House series of Intelligence tests, English tests, and Arithmetic tests.

Since the main body of data used in the present investigation comprises tests of ten items each we shall use as an example a ten item test. Suppose that an examiner wishes to construct a ten item test producing a score-scatter specified by its mean, its standard deviation, and its skewness. The mean of the corresponding answer-pattern-differential is then fixed as equal to the mean of the intended score-scatter. The standard deviation and the skewness of the answer-pattern-differential to be used must be calculated from the regression equations of chapters 6 and 7, or, more easily, read off from the regression lines.

There are thus three quantities given, sufficient to fix only three points of the answer-pattern. The first stage must then be the construction of a three point answer-pattern, such as would be obtained from a three item test, with a mean three-tenths of the

given mean, a standard deviation diminished in the same ratio, but an unaltered skewness, since that is already measured in standardised units.

Let the ordinates of this three item answer-pattern be $n_0$, $n_1$, $n_2$, $n_3$; the answer-pattern -differential is therefore $n_0-n_1$, $n_1-n_2$, $n_2-n_3$, $n_3$. Let $y_0 = \frac{n_0 - n_1}{n_0}$ , $y_1 = \frac{n_1 - n_2}{n_0}$ , and so on. Then

$$y_0 + y_1 + y_2 + y_3 = 1 .$$

If the first, second, and third moments of this y distribution about zero are denoted by $m_1$, $m_2$, and $m_3$, then

$$\begin{aligned} y_1 + 2y_2 + 3y_3 &= m_1 \\ y_1 + 4y_2 + 9y_3 &= m_2 \\ y_1 + 8y_2 + 27y_3 &= m_3 \end{aligned}$$

Now it is easy to prove that $m_1 = a$, $m_2 = \sigma^2 + a^2$, $m_3 = \sigma^3 S + 3\sigma^2 a + a^3$ , where $a$, $\sigma$, and $S$ are the mean, standard deviation, and skewness of the three item answer-pattern-differential.

Thus we obtain the four equations

$$\begin{aligned} y_0 + y_0 + y_2 + y_3 &= 1 \\ y_1 + 2y_2 + 3y_3 &= m_1 = a \\ y_1 + 4y_2 + 9y_3 &= m_2 = \sigma^2 + a^2 \\ y_1 + 8y_2 + 27y_3 &= m_3 = \sigma^3 S + 3\sigma^2 a + a^3. \end{aligned}$$

From these equations the values of $y_0$, $y_1$, $y_2$, $y_3$ may be calculated. The solution is made easier if the m's are

first calculated from the given values of a, $\sigma$, and S. Then the required values of the y's interms of the known m's are;

$$y_0 = 1 - \tfrac{11}{6}m_1 + m_2 - \tfrac{1}{6}m_3$$
$$y_1 = 3m_1 - \tfrac{5}{2}m_2 + \tfrac{1}{2}m_3$$
$$y_2 = -\tfrac{3}{2}m_1 + 2m_2 - \tfrac{1}{2}m_3$$
$$y_3 = \tfrac{1}{3}m_1 - \tfrac{1}{2}m_2 + \tfrac{1}{6}m_3$$

From these values of y the values of n giving the required answer-pattern are easily calculated for any given number of candidates. An example will make the application of the method clearer, and will serve as a basis for the discussion of the processes involved in converting this three item pattern into a ten item pattern.

2. An example.

It is desired to construct a ten item answer-pattern giving an answer-pattern-differential with mean score 4, standard deviation 4, and skewness +0.5 .

The corresponding values for the three item answer-pattern-differential are a = 1.2, $\sigma$ = 1.2, S = +0.5 .

Hence $m_1 = 1.2$, $m_2 = 2.88$, $m_3 = 7.77$ .

Hence $y_0 = 0.38$, $y_1 = 0.29$, $y_2 = 0.07$, $y_3 = 0.26$.

For 100 candidates this would mean an answer-pattern-differential 38, 29, 7, 26; that is, an answer-pattern 100, 62, 33, 26.

It is profitable for us to analyse a little more fully some of the steps in the above calculation, to bring out their

implications. When the mean was fixed at 1.2, $m_1$ was thereby fixed at 1.2, and the values of $m_2$ and $m_3$ were also partially fixed, since they are functions of the mean and other variables. Similarly fixing the standard deviation at 1.2 finally fixed the value of $m_2$, and still further circumscribed the range of possible values of $m_3$, or in other words delimited the range of possible values of S. The limitation of the range of possible values of S arises through each y being a positive (or zero) quantity not more than unity.

If, in the example, we substitute the values for a and $\sigma$, and then evaluate $y_0$, $y_1$, $y_2$, $y_3$, we can find the limits within which S must lie. The equations become

$$y_0 = .525 - .288S$$
$$y_1 = -.144 + .864S$$
$$y_2 = .504 - .864S$$
$$y_3 = .112 + .288S$$

The limits of each y are 0 and 1, by the definition of answer-patterns. Hence we deduce from the respective equations the following pairs of rough limits of S.

$+2 > S > -2 : +0.2 < S < +1 : +0.6 > S > -0.6 : -0.4 < S < +3$ .

Taken together the four equations limit the permissible values of S to the range +0.2 to +0.6. The value actually chosen was +0.5 .

The process might have been reversed. S might have been first fixed, then $\sigma$, and finally the mean a would have had to be

chosen from a restricted range. The three variables concerned may be inserted in any order.

The next step is the conversion of this three item pattern to a ten item pattern. In this process there is lacking that exactness of method and fixity of results which characterized the preceding part of the calculation. There are many possible ways of filling out a three item answer-pattern to one of ten items. Of these, there are two that merit further investigation. They form the extremes, as regards steepness, of the possible methods.

The first is to give each of the first three items of the ten item test the same difficulty as the first item of the three item test; items four to seven the difficulty of the second item of the three item test; and items eight to ten the difficulty of the last item of the small test. In the example considered, the ten item pattern would be

100, 62, 62, 62, 33, 33, 33, 33, 26, 26, 26.

This gives an answer-pattern-differential

38, 0, 0, 29, 0, 0, 0, 7, 0, 0, 26,

with a mean 3.96, standard deviation 4.05, and skewness +0,55. The steepness of this test as measured by the coefficient c is 0.5 . It is the flattest test that can be constructed to fulfil the required conditions.

A second method of filling out the answer-pattern is to use the given values as representative points on a smooth curve, and

from this curve determine the values of the n's required. The representative points are placed at x = 2, $5\frac{1}{2}$, 9. The curve for the test considered is on page 146. From it we derive the answer-pattern

100, 78, 62, 51, 43, 36, 31, 29, 27, 26, 25,

which gives an answer-pattern-differential

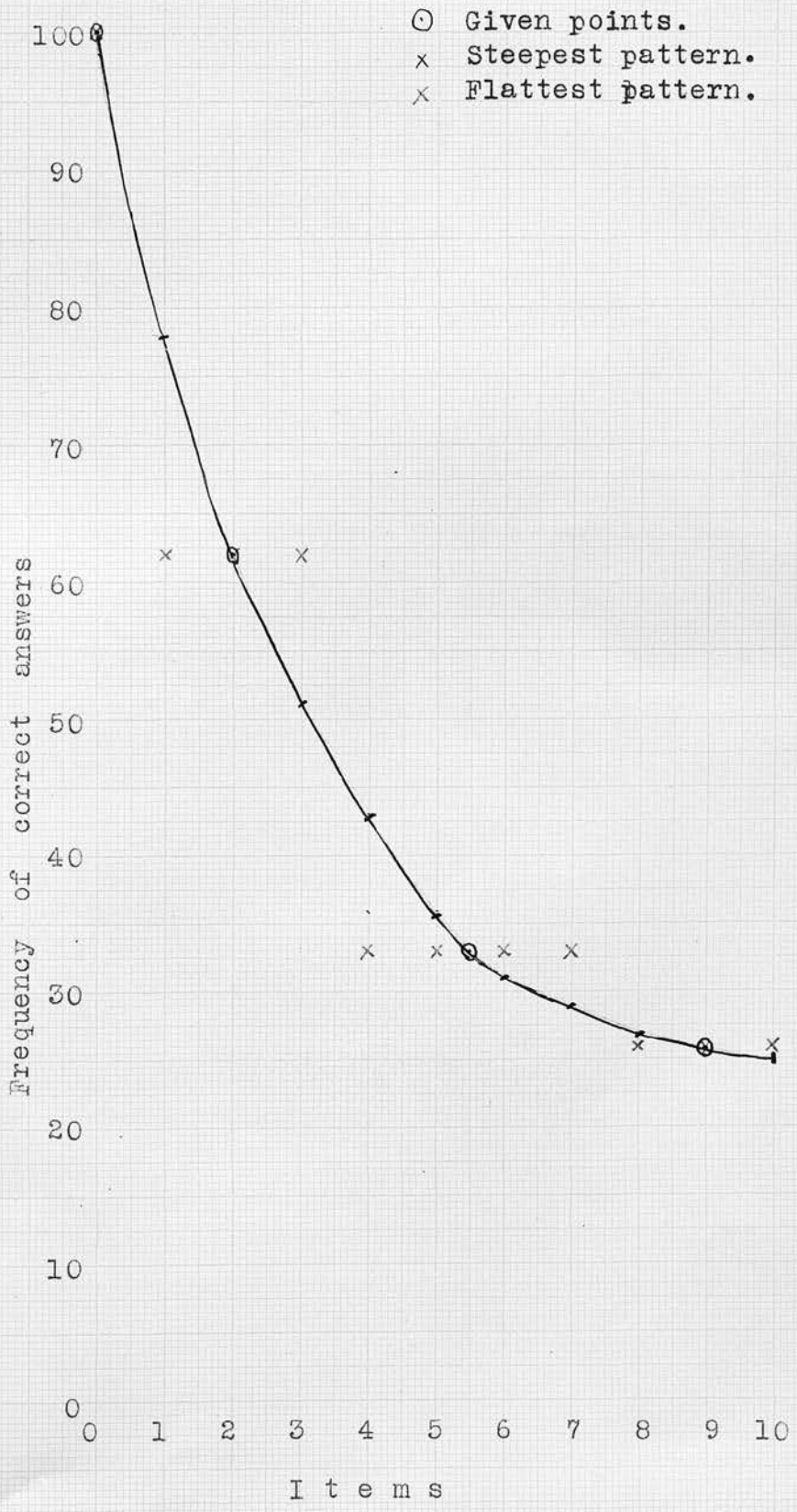
22, 16, 11, 8, 7, 5, 3, 1, 1, 1, 25,

with mean 4.07, standard deviation 3.89, and skewness +0.40. This is the steepest answer-pattern that can be constructed to fulfil the required conditions; its coefficient of steepness c equals 0.8 .

The tests which might be constructed to have the required mean, standard deviation, and skewness, are therefore restricted in steepness to the range 0.5 to 0.8 . As far as we know, there is no reason to prefer the steeper or the flatter type of test, save that a preference might be shown for the steeper test on the grounds mentioned on page 139. The final choice of the examiner would probably be determined by the items available.

If a test of 100 items were required, it would be necessary to amke use of regression equations connecting standard deviation of answer-pattern-differential and that of score-scatter, and connecting the coefficients of skewness of these distributions. It has been shown in chapter 7 that the equation for skewness is practically the same for 100 item tests as for 10 item tests. In the case of standard deviations this is not the case, probably

Fig. 25. Completion of answer-pattern from given points.



because of the lack of "headroom" in the 10 item tests. It would be necessary first to collect sufficient data to construct a reliable regression equation for 100 item tests. This the author has been unable to do yet.

## Chapter 11. The Relation of the Reliability of Tests to the Nature of the Answer-pattern.

### 1. Theoretical relations of reliability and hig.

The reliability of tests of the type under consideration is usually measured by the split-halves method. If the correlation between the scores on the odd items and those on the even items is denoted by $r_{\frac{1}{2}}$, then the coefficient of reliability is defined as $\frac{2r_{\frac{1}{2}}}{1 + r_{\frac{1}{2}}}$. It represents the correlation that would exist between the scores made in the complete test and those made in a similar test if such existed.

There are obviously some points of connection between this coefficient and the answer-pattern of the test. It is interesting to note that the present investigation arose from a discussion among certain educationists, one of whom claimed that the reliability of tests such as group tests of intelligence was a fictitious reliability, its magnitude being largely preordained by the tester eliminating as far as possible differences due to hig. The author was asked to investigate this problem, which soon resolved itself into a much wider series of problems, and the fruits of the research are submitted in this thesis. The original problem may now appear to be rather a bypath from the main track, but it is quite important from the point of view of those constructing and using tests of the type discussed here.

Before the reliability coefficient is calculated, the items

of the test should be placed in order of difficulty. Under these conditions it is possible to prove that unig is sufficient but not necessary, to produce perfect reliability.

(1) Unig implies perfect reliability.

If the answers to the test are unig each score 2x is made up of answers to the x easiest odd items and the x easiest even items. Thus there is perfect correlation between scores on odd and even items. Scores which are odd, and therefore cannot be halved exactly, will produce small variations which may in practice depress the correlation slightly below the theoretical maximum.

(2) Unig is not necessary for perfect reliability.

This is a negative proposition which is most easily proved by constructing a particular case, a test which has perfect reliability and yet some degree of hig. Such a test may be constructed as under.

Consider a 6 item test, the items being numbered 1,3,5,7,9,11 Suppose that the following results are obtained when the test is attampted by 10 candidates A - J.

(Table 33)

Table 33. Results of a 6 item test.

| Candidates | Items 1 | 3 | 5 | 7 | 9 | 11 | Scores |
|---|---|---|---|---|---|---|---|
| A | | | x | | | | 1 |
| B | | x | | x | | | 2 |
| C | x | | | | x | | 2 |
| D | x | | x | | x | | 3 |
| E | x | x | x | | | | 3 |
| F | x | x | x | | | | 3 |
| G | x | x | x | | | | 3 |
| H | x | x | x | x | | | 4 |
| I | x | x | | x | | x | 4 |
| J | x | x | | x | x | x | 5 |
| Answer-pattern | 8 | 7 | 6 | 4 | 3 | 2 | 30 |

| Score-scatter | x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|
| | $N_x$ | 0 | 1 | 2 | 4 | 2 | 1 | 0 |

It is apparent from the table that a certain amount of hig is present; h, as defined in chapter 8, equals .28 .

This test could be converted into a 12 item test of perfect reliability merely by doubling the table, making item 2 a duplicate of item 1 and so on. This would not be a satisfactory method pf proof, as the resulting test would be of a most unusual type. In any case such a procedure is quite unnecessary, as it is quite easy to construct another test with the same answer-pattern

and score-scatter. Such a test is shown below, the items being numbered 2,4,6,8,10,12. Again h = .28 .

Table 34. Results of a 6 item test.

| Candidates | Items 2 | 4 | 6 | 8 | 10 | 12 | Scores |
|---|---|---|---|---|---|---|---|
| A | x | | | | | | 1 |
| B | | | x | | x | | 2 |
| C | x | x | | | | | 2 |
| D | x | x | | | | x | 3 |
| E | x | x | x | | | | 3 |
| F | x | | x | x | | | 3 |
| G | x | x | x | | | | 3 |
| H | x | x | | x | x | | 4 |
| I | x | x | x | x | | | 4 |
| J | | x | x | x | x | x | 5 |
| | 8 | 7 | 6 | 4 | 3 | 2 | 30 |
| | Answer-pattern | | | | | | |

| Score-scatter | x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|
| | $N_x$ | 0 | 1 | 2 | 4 | 2 | 1 | 0 |

If the results of these two tests are now combined to give a 12 item test, the results are as shown in the following table. It will be observed that each candidate's score may be equally divided into answers to odd and even items. That is, the reliability of the test is perfect.

Table 35. Combined results giving a 12 item test.

| Candidates | Items 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Scores |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | x | | | x | | | | | | | | 2 |
| B | | | x | | | x | x | | | x | | | 4 |
| C | x | x | | x | | | | | x | | | | 4 |
| D | x | x | | x | x | | | | x | | | x | 6 |
| E | x | x | x | x | x | x | | | | | | | 6 |
| F | x | x | x | | x | x | | x | | | | | 6 |
| G | x | x | x | x | x | x | | | | | | | 6 |
| H | x | x | x | x | x | | x | x | | x | | | 8 |
| I | x | x | x | x | | x | x | x | | | x | | 8 |
| J | x | | x | x | | x | x | x | x | x | x | x | 10 |
| | 8 | 8 | 7 | 7 | 6 | 6 | 4 | 4 | 3 | 3 | 2 | 2 | 60 |

Answer-pattern

| Score-scatter | x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N_x$ | | 0 | 0 | 1 | 0 | 2 | 0 | 4 | 0 | 2 | 0 | 1 | 0 | 0 |

It might be objected that this is a very artificially constructed test, and that it is most unlikely that any test could be so neatly split up into the two components here shown. The reply to this objection is that it is merely another way of saying that a test with perfect reliability is unlikely to occur at all. What has been proved is that, if it did occur, there is no necessity for the test to be unig. It seems that the test is at least just as likely to show hig as to be unig.

When the items of the test are not in order of difficulty, unig is neither necessary nor sufficient to produce perfect

reliability. That unig is not necessary may be proved directly by suitable rearrangement of the items of the above test, care being taken to keep the odd items as odd, or to change all the odd items to even items, and vice versa. There will then be formed a test of perfect reliability which has a certain amount of hig in the answers.

A general argument may be used to establish the insufficiency of unig for perfect reliability when the items are not in order of difficulty. When the order of items is random, it is unlikely that any score of 2x, though made up of answers to the 2x easiest items, should be made up of answers to x odd and x even items, or that any more complicated relation between scores on odd and even items should exist causing perfect correlation. There are m! - 1 ways of rearranging the m items of a unig test. Of these $\left\{\left(\frac{m}{2}\right)!\right\}^2 - 1$ preserve odds as odds and evens as evens. The probability of the reliability being still perfect after rearrangement is therefore $\left(\left\{\left(\frac{m}{2}\right)!\right\}^2 - 1\right)/(m! - 1)$, which for m= 10 roughly equals 1 in 252.

The sole positive conclusion that has been drawn, then, is that unig type of answering implies perfect reliability when the items are in order of difficulty. This may mean that tests in which the incidence of hig has been reduced will have on that account a higher reliability coefficient, but such a conclusion must be established by evidence from experiments.

2. Experimental evidence on the relation of reliability and hig.

An initial difficulty here is that no really satisfactory measure of the quantity of hig in a test has yet been devised. There have been described the coefficient of hig, h; the coefficient $\alpha$ measuring the ratio of the influence of the answer-pattern-differential to that of a hypothetical normal distribution; and the coefficient of steepness, c, which may bear some relation to the amount of hig present. These must serve meantime.

(a) Data of 41 tests.

In the case of the 41 tests all these coefficients are available. The reliability of one of those tests must be calculated by placing the items in order of difficulty, so classifying the items as odd or even, and then correlating the scores on odd and even items. These scores must, of course, be obtained from the original data.

Each of the 41 tests is a 10 item test, so that the scores to be correlated are those on the 5 odd items with those on the 5 even items. This naturally leads to a coarseness of grouping effect in the correlation table, with an adverse effect on the accuracy of the correlation coefficient. Also the number of candidates is small for statistical purposes, so that the coefficients obtained have rather high probable errors. For these reasons the reliabilities only of certain selected tests were calculated; these were the tests showing the least and greatest values of h, $\alpha$, and c. The results were as follows.

Table 36. Relation of reliability to h, $\alpha$, and c.

| Test | Special feature | h | $\alpha$ | c | Reliability |
|---|---|---|---|---|---|
| 27 | Smallest h | .06 | .31 | .9 | .82 |
| 41 | Greatest h | .41 | .42 | .8 | .77 |
| 39 | Greatest $\alpha$ | .19 | .50 | .9 | .70 |
| 24 | Smallest $\alpha$ | .20 | -.22 | .8 | .49 |
| 20 | Greatest c | .14 | .04 | .9 | .71 |
| 27 | | .06 | .31 | .9 | .82 |
| 34 | | .11 | .33 | .9 | .74 |
| 39 | | .19 | .50 | .9 | .70 |
| 11 | Smallest c | .34 | .23 | .6 | .81 |
| 28 | | .40 | .20 | .6 | .79 |
| 32 | | .33 | -.17 | .6 | .73 |
| 33 | | .26 | -.10 | .6 | .73 |

It is evident from the above table that there is no clear relationship between reliability and hig as measured by h, $\alpha$, or c.

(b) Data of physics tests.

In the case of this and other groups of complete tests, no values of $\alpha$ are available. The difficulty of coarseness of grouping in the correlation table is even more pronounced with these physics tests, since they were 8 item tests. It was probably on this account that some of the reliabilities obtained were so low. In one test, (F), perfect scores obtained by 4 candidates were omitted. If allowed to stand these scores would increase the reliability coefficients in an artificial way. It must also be noted that the coefficient of steepness given is calculated from a test of 8 items, and may be used only to compare the steepnesses of 8 item tests, as is done in the table.

Table 37. Relation of reliability to h and c, tests D, E, F.

| Test | $n_o$ | h | c | Reliability |
|---|---|---|---|---|
| D | 34 | .18 | .88 | .19 |
| E | 34 | .07 | .67 | .49 |
| F | 30 | .22 | .81 | .56 |

Once again the results are not very enlightening.

(c) Data of thesis tests.

Tests A, B, and C provide much more suitable material. Each test contained 15 items, so that the effects of coarseness of grouping were not so evident. The number of candidates was also reasonably large. On this occasion it was zero scores which had to be eliminated from the data, to avoid an artificial boosting of the correlation. The pmission of these scores does not affect the value of h, but the coefficient c must be recalculated, since the path from $n_o$ to $n_1$ has been altered, and with it all three lengths considered in the definition of that coefficient When these precautions had been taken the results were as follows.

Table 38. Relation of feliability to h and c, tests A, B, and C.

| Test | $n_o$ | h | c | Reliability |
|---|---|---|---|---|
| A | 118 | .40 | .63 | .77 |
| B | 159 | .09 | .76 | .75 |
| C | 160 | .07 | .86 | .86 |

The reliability of the steepest test (C) is significantly greater than that of either of the others, when the usual test is applied.

Unfortunately no reliability coefficients seem to be available for the M.H.T. group of tests. Perhaps the loss is more apparent than real, for these tests are nearly all of the steep type for the number of items they contain. This would greatly diminish their usefulness as a group in which to study the relation of hig to reliability. The only test to which this does not apply, test 12p, is also rather spoiled for comparison with the others, for it is of the flat type, which tends to increase the incidence of hig, but it has very few items, which tends to decrease the incidence of hig. Whether the one effect compensates the other we cannot say.

On the experimental evidence considered above it is impossible to decide whether the minimising of hig in a test thereby increases its reliability. The solution of this problem, as of others raised in this thesis, must await the advent of more extensive data.

PART TWO.

Notes on the Moray House Tests of Intelligence referred to in Part One, with tables of data.

These notes are intended to indicate features of interest in the Moray House series of Tests from the point of view of the preceding chapters. Through the kindness of Professor G. H. Thomson, there is included in this part a copy of each of the tests. Following each are tables of the frequencies of correct answers and the score-scatters obtained when the test was applied to specified populations.

There are points of interest common to all the tests. One is that the items have been arranged roughly in increasing order of difficulty. This fact, and the direction printed on most "Begin at the beginning, and go straight through" tend to reduce the amount of hig in all the tests. That amount, as measured by h, is low in all. A second point is that in most of the tests the answer-pattern is a straight line sloping from a very easy item to a very difficult item. With such an answer-pattern, and unig type of answering, the score-scatter would show the same number of candidates for every score in the range. Now the score-scatters produced are almost normal, or Gaussian, distributions. It follows that this normality is not caused by, but rather occurs in spite of, the nature of the answer-pattern coupled with a low degree of hig.

The tests included below are M.H.T. 8, 9, 11, 12v and 12p.

M. H. T. 8.

# DO NOT OPEN THIS BOOK UNTIL YOU ARE TOLD.

Examination Number only.

# LANCASHIRE EDUCATION COMMITTEE.

## EXAMINATION FOR JUNIOR SCHOLARSHIPS, 14th FEBRUARY, 1931.

## INTELLIGENCE TEST.

Age last Birthday.....

Date of Birthday.....

## INSTRUCTIONS.

When you are told to begin, answer the questions as quickly and as carefully as you can.

Begin at the beginning and go straight through.

If you cannot do any question in any test, leave it out and go on to the next.

When you finish one page, go on to the next.

You will have 45 minutes, and you will be told the time every quarter of an hour. No one is expected to do everything. Just do as much as you can.

ASK NO QUESTIONS AT ALL.

## TEST 1a.—FOLLOWING DIRECTIONS.

**Read each question carefully, and then write the answer to it in the bracket.**

**The alphabet is printed here to help you :—**

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

**BEGIN HERE :—**

(1) Do you understand you must not look at the paper of any other pupil during this examination ? If so, write P ... ... ... ... ... ... ... ( )

(2) Write the letter which comes before R in the alphabet ... ... ... ... ( )

(3) Write the odd numbers between 1 and 9, and underline the middle one ( )

(4) If 22 farthings are the same as $6\frac{1}{2}$ pence, write O ; if not, write C ... ... ... ( )

(5) Write the letter which occurs most often in the word IRRADIATION ... ... ( )

(6) If M and K are separated in the alphabet by one letter, write it down ; if not, write A ... ... ... ... ... ... ... ... ... ... ... ( )

(7) If M comes after R in the alphabet, write P ; if not, write X ... ... ... ( )

(8) If $\frac{1}{8}$ is more than $\frac{1}{7}$, write N ; if not, write Y ... ... ... ... ... ... ( )

## TEST 2a.—ANALOGIES.

Look at the first example :—

(1) finger : hand—toe : ? ... ... (foot, knee, arm, shoe, nail)

This means that finger is to hand as toe is to what ? The answer is one of the five words in the bracket. the right answer is " foot " and it is underlined, since a finger is part of a hand just as a toe is part of a foot.

Now look at Example (2) :—

(2) man : clothes— ? : fur ... ... (coat, animal, bird, skin, cloth)

This means that man is to clothes as what is to fur ? " Animal " is the right answer, because an animal wears fur just as a man wears clothes ; so " animal " is underlined.

Now look at Example (3) :—

(3) king : queen—lord : ? ... ... (princess, sister, duke, lady, prince)

In each line you are to look at the five words in the bracket, and decide which of them should go where the question mark is, and underline it. Do them just as the examples were done. Remember, you have nothing to do but to UNDERLINE ONE word in each bracket.

**BEGIN HERE :—**

(1) brother : sister—nephew : ? ... (cousin, niece, boy, girl)
(2) up : down—west : ? ... ... (north, opposite, east, over, south)
(3) ship : steamer— ? : tiger ... (animal, eagle, lion, camel, runner)
(4) king : country— ? : school ... (teacher, caretaker, scholar, prince, headmaster)
(5) adjective : noun— ? : verb ... (proverb, adverb, subject, object, preposition)
(6) lean : fat—small : ? ... ... (full, many, empty, large, much)
(7) field : gate—house : ? ... ... (window, room, chimney, door, wall)

## TEST 3a.—NUMBER SERIES.

Look at the first line of numbers :—

Example (1) 1 2 3 4 5 ... ... ... ... ... ... ... ( 6 )

The one that comes next is 6, because the numbers go up one at a time. In each line there is a rule for finding the next number. In this one the rule is that the numbers go up by 1 each time. The other lines have different rules.

Example (2) 12 10 8 6 4 ... ... ... ... ... ... ... ( 2 )

Here the rule is that the numbers come down by 2 at each time.

Example (3) 1 2 4 8 16 ... ... ... ... ... ... ... ( 32 )

Here the rule is that each number is twice as big as the one before it, so the answer in the bracket is 32.

Now try the lines below. In each line find the rule, and then write the number that should come next in the bracket.

**BEGIN HERE :—**

2 5 8 11 14 17 ... ... ... ... ... ... ... ... ( )

31 28 25 22 19 ... ... ... ... ... ... ... ... ... ( )

$\frac{1}{11}$ $\frac{1}{10}$ $\frac{1}{9}$ $\frac{1}{8}$ $\frac{1}{7}$ ... ... ... ... ... ... ... ... ... ( )

28 21 14 7 ... ... ... ... ... ... ... ... ... ... ( )

4 8 12 16 12 8 ... ... ... ... ... ... ... ... ( )

3 3 5 5 7 ... ... ... ... ... ... ... ... ... ( )

2 6 18 54 ... ... ... ... ... ... ... ... ... ... ( )

## TEST 4a.—REASONING.

DIRECTIONS.—Three answers to each question are given in the bracket after it. You are to underline what you think is the RIGHT answer. You have nothing to write. Only UNDERLINE.

(1) Tom has more money than Dick, and Dick has more money than Harry. Who has the most money of the three ? (Tom, Dick, Harry)

(2) Ada is smaller than Bertha, but not so small as Clara. Who is the smallest of the three ? ... ... ... (Ada, Bertha, Clara)

(3) Wool is dearer than cotton, and silk is dearer than wool. Which is the cheapest ?... ... ... ... ... (wool, silk, cotton)

(4) Mr. Smith's house is larger than Mr. Jack's, and Mr. Watt's is larger than Mr. Smith's. Who has the largest house ? ... ... ... ... ... ... ... (Mr. Smith, Mr. Jack, Mr. Watt)

(5) Three people A, B, C, set out from London. A goes half as far as C, but twice as far as B. Who went farthest ? (A, B, C)

(6) Mr. Ross's house is near the grocer's shop, but Mr. Page's house is nearer still ; while Mr. Robb's house lies between the other two. Who is nearest the grocer's shop ? (Mr. Ross, Mr. Page, Mr Robb)

## TEST 5a.—CLASSIFICATION.

Look at the first example :—

Example (1): bullet cannon gun pencil sword

Here we have a line of five words. One of them, "pencil," has been underlined. The other four words of names of things used for fighting. But a pencil is not used for fighting, so we underline it.

Now look at the second example :—

Example (2): grass bread meat milk potatoes

Again, we have a line of five words. "bread," "meat," "milk," and "potatoes" are the names of things we eat. But we do not eat "grass," so we underline it.

Look at the third example :—

Example (3): mill fill pill spill say

Again, we have a line of five words. The first four—"mill," "fill," "pill," "spill," sound like each other; but "say" does not sound like them at all. It sounds quite different, so we underline it.

Now try the following. In each line underline JUST ONE WORD that does not belong there.

| | | | | | |
|---|---|---|---|---|---|
| (1) | red | square | blue | green | yellow |
| (2) | ball | cube | circle | coin | ring |
| (3) | wood | lead | iron | copper | gold |
| (4) | Mary | Tom | Sam | Dick | John |
| (5) | bad | grand | fine | splendid | good |
| (6) | arm | skin | hair | glove | foot |
| (7) | big | huge | small | large | great |
| (8) | begin | originate | commence | start | finish |

## TEST 1b.—FOLLOWING DIRECTIONS.

**Read each question carefully, and then write the answer to it in the bracket.**

**The alphabet is printed here to help you :—**

**A B C D E F G H I J K L M N O P Q R S T U V W X Y Z**

(1) If X comes after V in the alphabet, and if L comes after P, write G; but if only one of these is true, write M ... ... ... ... ... ... ... ... ( )

(2) The first letters of the four directions (East, etc.) when put in a certain order form a word; write it in the bracket ... ... ... ... ... ( )

(3) HTOMMAM is a word seen in a mirror. Write it as it usually appears ( )

(4) Write P, unless the second letter of this sentence is R; if it is, write T ... ... ( )

(5) If the letters in the alphabet were written starting from the other end, what would the 13th letter be? ... ... ... ... ... ... ... ... ... ( )

(6) Suppose all the even letters in the alphabet came first, then the odd ones, what would the fifth letter of the alphabet then be? ... ... ... ... ( )

(7) Write the letter which follows the letter which comes after E ... ... ... ( )

(8) If the alphabet began at K, the preceding letters being put at the end, what would the 8th letter be? ... ... ... ... ... ... ... ... ... ( )

(9) Write the letter which is next but one after the letter between K and M ... ... ( )

## TEST 2b.—ANALOGIES.

This is like Test 2a on page 2. You may look back at the directions if you wish.

You underline ONE word in each bracket.

(1) hat : head—glove : ? ... ... ... (arm, hand, elbow, foot, face)

(2) A : Z—beginning : ? ... ... ... (after, start, complete, end, distant)

(3) needle : prick—knife : ? ... ... (sharp, fork, point, blade, cut)

(4) hive : bee— ? : man ... ... (food, work, honey, meat, house)

(5) a : d — first : ? ... ... (second, fourth, later, last, third)

(6) pleasure : rejoice—doubt : ? ... ... (sorrow, lament, believe, act, hesitate)

(7) cellar : coal— ? : milk ... ... (man, tea, coffee, tub, jug)

(8) anger : pleasure—rage : ? ... ... (hesitation, delight, sorrow, lament, expect)

## TEST 3b.—NUMBER SERIES.

This is like Test 3a on page 3. You may look back at the directions if you wish.

Write the number that comes next in the bracket.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\frac{1}{2}$ | $\frac{1}{6}$ | $\frac{1}{8}$ | $\frac{1}{11}$ | $\frac{1}{14}$ | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 3 | $4\frac{1}{2}$ | 6 | $7\frac{1}{2}$ | 9 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| $5\frac{1}{4}$ | $8\frac{1}{4}$ | $11\frac{1}{4}$ | $14\frac{1}{4}$ | $17\frac{1}{4}$ | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 4 | 6 | 9 | 13 | 18 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 2 | 6 | 12 | 20 | 30 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 1 | 4 | 16 | 64 | 256 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 20 | 19 | 17 | 14 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |

## TEST 4b.—REASONING.

**Remember that you underline the RIGHT answer.**

(1) John is better than Tom at composition. But Tom is better than John at drawing. If composition is more important than drawing, who is the better of the two ? (John, Tom, No one can tell)

(2) Tom, John and Peter are sitting in a row. John is on the left of Peter, and Tom is to the left of John. Who is in the middle ? ... ... ... ... ... (John, Tom, Peter)

(3) If Dick is put on the right of Harry, and Andrew is put on the right of Dick, who is now in the middle ? ... (Dick, Andrew, Harry)

(4) All kinds of wood float on water. A piece of material is thrown on water, and it floats. Is it wood ? ... (Yes, No, No one can tell)

(5) "Umo" is twice as dear as "Ritka"; "Ritka" is twice as good for food as "Umo" is. I have 1/- to spend. Is it better to buy "Umo" or "Ritka" ? ... ("Umo", "Ritka", No one can tell)

**Go on to NEXT PAGE without waiting to be told.**

## TEST 5b.—CLASSIFICATION.

Remember that in each line there is one word that does not belong there ; it is different in some way from the others.

When you have found it, underline it. Underline JUST ONE WORD in each line.

| | | | | | |
|---|---|---|---|---|---|
| (1) | jump | leap | halt | walk | run |
| (2) | receive | give | find | beg | borrow |
| (3) | cloth | butter | bread | cake | beef |
| (4) | hurt | pleasure | enjoyment | gladness | ease |
| (5) | mend | create | construct | destroy | produce |
| (6) | Cæsar | Napoleon | Milton | Haig | Wellington |
| (7) | pencil | ink | pen | chalk | crayon |
| (8) | gramophone | violin | sonata | piano | organ |
| (9) | assemble | gather | amass | collect | disperse |

## TEST 1c.—FOLLOWING DIRECTIONS.

**Remember you write the answer to the question in the bracket.**

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

(1) If the word FACETIOUS contains the vowels in their proper order, write L ; if not, write P ... ... ... ... ... ... ... ... ... ... ( )

(2) Write the two letters in the word SMALLER that have as many letters between them in the alphabet as there are letters in the word PORT ... ... ( )

(3) Write X if SUBSEQUENT contains the 17th letter of the alphabet, unless F comes in the alphabet after the second letter of the word CHART, in which case write C ... ... ... ... ... ... ... ... ... ... ( )

(4) If the letters in the word BEGINS appear in the order in which they are found in the alphabet, write the letter which is midway in the alphabet between the second and sixth letters in the word ; if not, write O ... ... ... ( )

(5) Harry said over the letters of the alphabet till he came to M ; then went backwards for six letters ; then forward again for two letters. What was the letter to the left of that at which he stopped ? ... ... ... ... ... ... ( )

(6) If the letters in the word YES appear in the same order as they do in the alphabet ; and if the same is true for the letters of the word NO, write X ; but if this is true of only one of the words, write T ... ... ... ... ... ... ( )

(7) If each pair of letters in the alphabet were interchanged, so that it now read BADC . . . . , what would the 13th letter in the alphabet be then ? ... ( )

(8) Write the letter in the alphabet midway between the two letters which occur most often in the word VICISSITUDES ... ... ... ... ... ... ( )

## TEST 2c.—ANALOGIES.

**Remember you underline the right word in each bracket.**

(1) air : breathing—water : ? ... (swimming, floating, washing, drinking)
(2) calculate : reason—anger : ? ... (thought, feeling, imagination, memory)
(3) hand : foot—finger : ? ... ... (leg, ankle, toe, palm, elbow)
(4) petrol : car— ? : train ... ... (wheels, smoke, engine, steam)
(5) popular : applause—criminal : ?... (punishment, misery, reward, judgment

**Go on to NEXT PAGE without waiting to be told.**

## TEST 3c.—NUMBER SERIES.

**Remember that you write the number that should come next in the bracket.**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 6 | 12 | 20 | 30 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 5 | 10 | 20 | 40 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 1 | 4 | 2 | 5 | 3 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 2 | 7 | 14 | 23 | 34 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 2 | 3 | 2 | 4 | 2 | 5 | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 7 | 1 | 6 | 2 | 5 | 3 | 4 | ... | ... | ... | ... | ... | ... | ( ) |
| 600 | 300 | 100 | 25 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |
| 625 | 125 | 25 | 5 | 1 | ... | ... | ... | ... | ... | ... | ... | ... | ( ) |

## TEST 4c.—REASONING.

**Remember that you underline the right answer.**

(1) As I stand with my back to the rising sun, my house is on my left hand. In what direction must I walk to get home ? (North, South, West)

(2) A poet writes three poems. The first has three verses of four lines each ; the second has two verses of six lines each ; the third has five verses of two lines each. Which is the shortest poem ; ... ... ... ... ... (first, second, third)

(3) Three cars travel along a road. A passes B but cannot pass C. The one now at the back increases its speed, and passes the other two. Which is now farthest behind ? ... ... ... ... ... ... ( A, B, C )

(4) There is a town, all of whose streets run either North and South, or East and West. Walking along Brown Street, I am going East. I turn to the left along Hillside Street ; then to the right along London Street, then to the left along Oxford Street. In what direction does London Street run ? ... ... ... ... ... ... (North and South<br>East and West<br>No one can tell)

(5) Mrs. Smith, Mrs. Brown and Mrs. Jones buy the same kind of cloth at the same shop. Mrs. Smith buys much more than Mrs. Brown, who, however, spends a little less than Mrs. Jones. Their daughters, Miss Smith, Miss Brown and Miss Jones buy a hat each, and the account paid by each mother for herself and her daughter is the same. Which of the daughters chose the most expensive hat ? ... ... ... ... (Miss Smith, Miss Brown, Miss Jones)

## TEST 5c.—CLASSIFICATION.

Remember that you underline the word in each line that does not belong there.

Cross out JUST ONE WORD in each line.

| | | | | | |
|---|---|---|---|---|---|
| (1) | crystal | wall | spectacles | bottle | window |
| (2) | telephone | tramcar | steamer | train | cab |
| (3) | Raphael | Collie | Spaniel | Terrier | Pomeranian |
| (4) | wheat | oats | turnips | rye | barley |
| (5) | bullet | knife | spoon | book | key |
| (6) | support | hinder | assist | help | encourage |
| (7) | red | violet | yellow | pansy | blue |
| (8) | explain | show | tell | describe | narrate |
| (9) | fairest | rarest | carest | farthest | greatest |

The frequencies of correct answers given below were obtained from the papers of 528 candidates, and the score-scatter from those of 6423 candidates. For these figures the author is indebted to the Director of Education of the county where this test was tried out. A first estimate of whether the 528 candidates are representative of the total 6423 can be made by comparing their average scpres, which were 71 for the sample and 74 for the whole group.

The frequencies of correct answers by the 528 candidates were given as percentages, and are tabulated as such. The answer-pattern and score-scatter are graphed after the tables.

In all these tests it may be taken for granted that the average age of the testees is eleven years.

Table 39. Percentages of correct answers in M.H.T. 8.

| | Subtests | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | 1a | 2a | 3a | 4a | 5a | 1b | 2b | 3b | 4b | 5b | 1c | 2c | 3c | 4c | 5c |
| 1 | 98 | 92 | 92 | 93 | 97 | 90 | 94 | 89 | 70 | 91 | 62 | 58 | 62 | 32 | 53 |
| 2 | 96 | 90 | 89 | 91 | 60 | 51 | 81 | 86 | 82 | 18 | 26 | 44 | 60 | 74 | 66 |
| 3 | 44 | 53 | 94 | 91 | 89 | 60 | 76 | 84 | 80 | 94 | 61 | 61 | 60 | 53 | 54 |
| 4 | 95 | 51 | 68 | 88 | 83 | 75 | 81 | 75 | 44 | 88 | 9 | 52 | 42 | 20 | 68 |
| 5 | 88 | 74 | 88 | 74 | 81 | 89 | 55 | 63 | 55 | 49 | 38 | 45 | 57 | 27 | 41 |
| 6 | 85 | 88 | 91 | 87 | 87 | 41 | 42 | 60 | - | 48 | 83 | - | 33 | - | 52 |
| 7 | 62 | 64 | 71 | - | 66 | 73 | 86 | 75 | - | 49 | 42 | - | 38 | - | 57 |
| 8 | 79 | - | - | - | 63 | 70 | 72 | - | - | 59 | 42 | - | 19 | - | 10 |
| 9 | - | - | - | - | - | 64 | - | - | - | 71 | - | - | - | - | 12 |

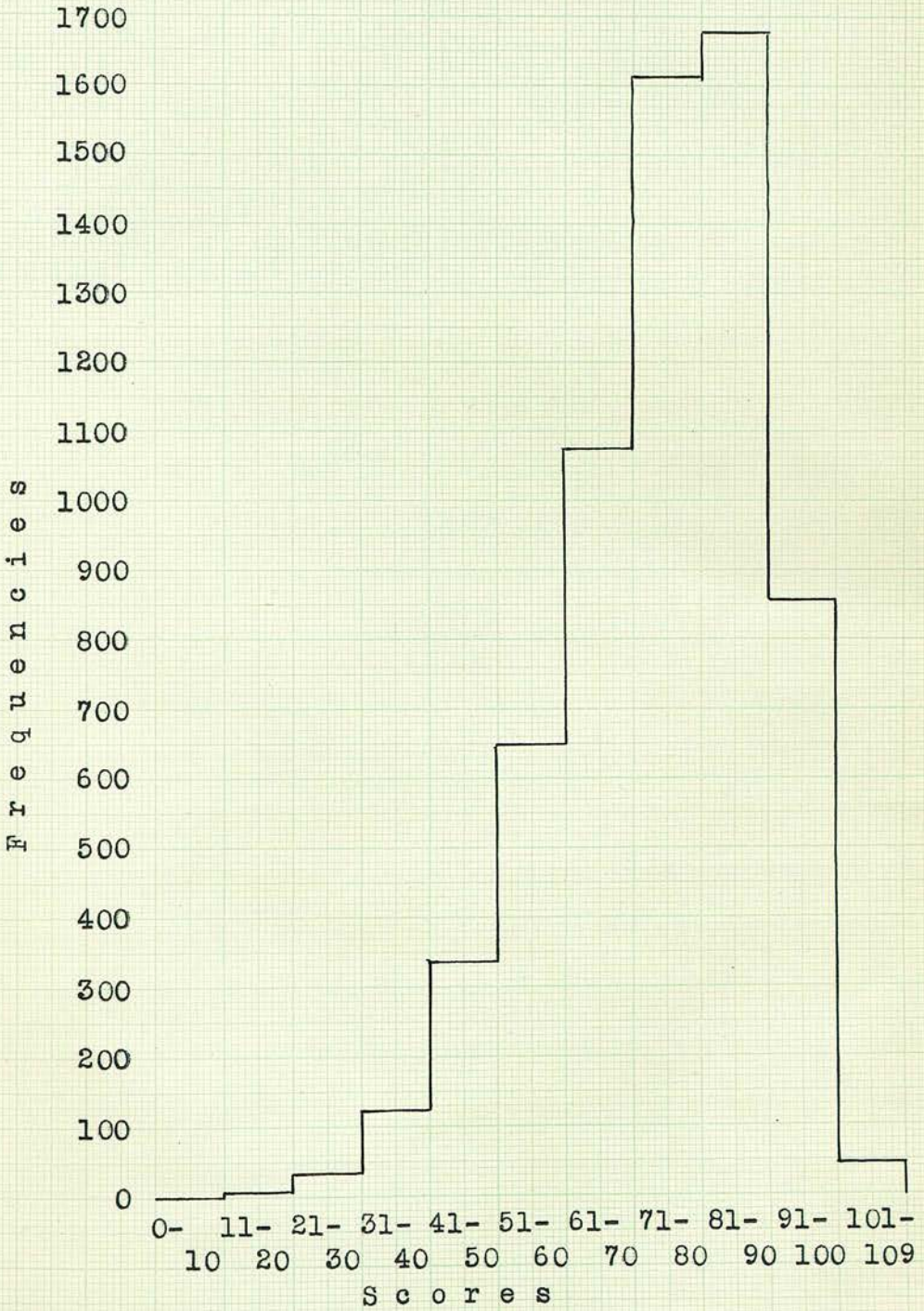
Table 40. Score-scatter of M.H.T. 8.

| Score | Frequency |
|---|---|
| 0 - 10 | 0 |
| 11 - 20 | 9 |
| 21 - 30 | 33 |
| 31 - 40 | 124 |
| 41 - 50 | 337 |
| 51 - 60 | 650 |
| 61 - 70 | 1076 |
| 71 - 80 | 1613 |
| 81 - 90 | 1679 |
| 91 -100 | 857 |
| 101 -109 | 45 |
| | 6423 |

This is a noteworthy case of the score-scatter being skewed negatively by a negatively skewed answer-pattern-differential. The skewness of the answer-pattern-differential is -0.38, and that of the score-scatter is -0.66 . The coefficient of hig is 0.11 .

Figure 26. Answer-pattern of M.H.T. 8.
( from 528 candidates )

Figure 27. Score-scatter of M.H.T. 8.
6423 candidates.



M. H. T. 9.

NOT TO BE FILLED IN BY THE SCHOLAR.

| Age (years and months) on 1/8/31. | Raw Score. | I.Q. |
|---|---|---|
| | | |

# County Borough of Halifax—Education Committee.

# INTELLIGENCE TEST, 1931.

Write your Surname here.....

Christian name here .....

School .....

What standard or class or form are you in ?.....

## INSTRUCTIONS.

When you are told to begin, answer the questions as quickly and as carefully as you can.

Begin at the beginning and go straight through.

If you cannot do any question in any test, leave it out and go on to the next.

When you finish one page, go on to the next. Be sure you do not turn over two pages at once.

You will have 45 minutes. You will be told the time every quarter of an hour.

*ASK NO QUESTIONS AT ALL.*

## TEST 1a.—FOLLOWING DIRECTIONS.

**Read each question carefully, then write the answer to it in the bracket.**

The alphabet is printed here to help you :—

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

**BEGIN HERE :—**

Do you understand that you must do your best and not ask questions ? If so, write M ( )

Have you noticed that there is an alphabet printed near the top of this page to help you ?
If so, write the 5th letter in the alphabet ... ... ... ... ... ... ( )

Write the letter before O in the alphabet ... ... ... ... ... ... ... ( )

Write the letter which occurs most often in the word UNDERSTANDING ... ... ( )

If the alphabet were written backwards, starting with Z, what would the 7th letter be ? ( )

Write the numbers between 4 and 9 and underline the smallest one ... ( )

If the letters K and L changed places in the alphabet, what would the 12th letter be ? ( )

If there are 14 farthings in $2\frac{1}{2}$ pence, write Y ; if not, write N ... ... ... ... ( )

## TEST 1b.—NUMBER SERIES.

Look at the first line of numbers :—

Example (1) 1 2 3 4 5 ... ... ... ... ... ... ... ( 6 )

The one that comes next is 6, because the numbers go up one at a time. In each line of numbers below there is a rule for finding the next number. In this one the rule is that the numbers go up by 1 each time. The other lines have different rules.

Example (2) 12 10 8 6 4 ... ... ... ... ... ... ... ( 2 )

Here the rule is that the numbers come down by 2 each time.

Example (3) 1 2 4 8 16 ... ... ... ... ... ... ... ( 32 )

Here the rule is that each number is twice as big as the number before it, so the answer in the bracket is 32.

Now try the lines below. In each line find the rule, and then write in the bracket the number that should come next.

2 4 6 8 10 12 ... ... ... ... ... ... ... ... ... ( )

3 7 11 15 19 ... ... ... ... ... ... ... ... ... ... ( )

2 2 3 3 4 ... ... ... ... ... ... ... ... ... ... ( )

1 $\frac{1}{3}$ $\frac{1}{5}$ $\frac{1}{7}$ $\frac{1}{9}$ ... ... ... ... ... ... ... ... ... ( )

16 13 10 7 4 ... ... ... ... ... ... ... ... ... ... ( )

2 6 18 54 162 ... ... ... ... ... ... ... ... ... ( )

## TEST 1c.—ANALOGIES.

Look at the first example :—

(1) finger : hand—toe : ? ... ... (foot, knee, arm, shoe, nail)

This means that finger is to hand as toe is to what ? The answer is one of the five words in the bracket. The right answer is "foot," so it is underlined ; it is the right answer, since a finger is a part of a hand, just as a toe is part of a foot.

Now look at Example (2) :—

(2) man : clothes— ? : fur ... ... (coat, animal, bird, skin, cloth)

This means that man is to clothes as what is to fur ? Now a man wears clothes just as an animal wears fur, so "animal" is the correct answer, and is therefore underlined.

Now look at Example (3) :—

(3) king : queen—lord : ? ... ... (princess, sister, duke, lady, prince)

In each line below you have to look at the five words in the bracket, decide which should go where the question mark is, and underline it. All you have to do is UNDERLINE ONE word in each bracket.

**BEGIN HERE :—**

horse : animal—swallow : ? ... ... (summer, fly, nest, bird, swift)
gate : field—door : ? ... ... ... (window, room, grass, stile, hinge)
feathers : hen—wool : ? ... ... (duck, jersey, sheep, blanket, coat)
arm : wrist—leg : ? ... ... ... (knee, elbow, ankle, bones, foot)
woman : girl— ? : boy ... ... (father, man, lad, youth, nurse)
long : short— ? : poor ... ... (thin, happy, small, content, rich)
before : after— ? : now ... ... (soon, then, but, why, if)
wing : bird— ? : fish ... ... ... (tail, mouth, fin, sea, feathers)

## TEST 1d.—REASONING.

DIRECTIONS.—Three answers to each question are given in the bracket after it. You are to underline what you think is the RIGHT answer. You have nothing to write. Only UNDERLINE ONE answer in each bracket.

**BEGIN HERE :—**

Segrave's car is faster than Campbell's car, and Campbell's car is faster than Harvey's. Who has the fastest car ? (Segrave, Campbell, Harvey)

Iron is stronger than wood, but not so strong as steel. Which is the strongest ? ... ... ... ... ... ... (iron, wood, steel)

John is taller than Tom, and Harry is taller than John. Who is the tallest ? ... ... ... ... ... ... (John, Tom, Harry)

Mary is older than Jane, and Ella is younger than Mary. Which of the three girls is the oldest ? ... ... (Mary, Jane, Ella)

I have three cricket bats. The first is heavier than the second, and the second is heavier than the third. Which is the lightest ? ... ... ... ... ... ... ... (first, second, third)

Three sticks of different lengths are coloured, one red, one blue, one green. The blue one is longer than the red one, and the green one is shortest of all. Of what colour is the longest stick ? ... ... ... ... ... ... (red, blue, green)

## TEST 1e.—LOGICAL SELECTION.

Look at this :—

dog ... (collar, hair, muzzle, chain, legs)

A dog ALWAYS has hair and legs, so these are underlined in the bracket. It does not always have a collar, or a muzzle, or a chain, so these are not underlined. In the lines below underline the TWO words which tell what the thing outside the bracket always has. Remember to UNDERLINE TWO words only in each bracket.

**BEGIN HERE :—**

boy ... (head, jacket, skin, hat, boots)

horse ... (stable, mouth, saddle, hoof, shoe)

motor car (petrol, smoke, engine, noise, wheels)

river ... (fish, banks, bridge, boat, water)

room ... (pictures, window, door, wall, table)

race ... (start, runners, spectators, competitors, prize)

## TEST 2a.—FOLLOWING DIRECTIONS.

**Read each question carefully, and then write the answer to it in the bracket.**

**The alphabet is printed here to help you :—**

**A B C D E F G H I J K L M N O P Q R S T U V W X Y Z**

If S is found before U in the alphabet, and M before K, write Y, but if only one of these is true, write C ... ... ... ... ... ... ... ... ... ... ( )

Four months of the year have names ending in the same three letters. Write the middle letter of these three ... ... ... ... ... ... ... ... ... ( )

Another four months have names ending in the same letter (not the same letter as in the previous question.) Write this letter ? ... ... ... ... ... ... ( )

If the alphabet had only 24 letters, F and G being dropped out, what would the 15th letter be ? ... ... ... ... ... ... ... ... ... ... ( )

OTTOM is a word seen in a mirror. Write it as it usually appears ... ( )

Write N unless the last letter of this sentence is R, in which case write that letter ... ( )

If the alphabet began with J, would the 10th letter be T ? If so, write A ; If not, write B ... ... ... ... ... ... ... ... ... ... ... ( )

If there are as many letters between H and M in the alphabet as there are between U and P, write Z ; if not, write X ... ... ... ... ... ... ... ( )

## TEST 2b.—NUMBER SERIES.

This is like Test 1b on page 2. You may look back at the directions if you wish. Write in the bracket the number that comes next.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 9 | 27 | 81 | ... | ... | ... | ... | ... | ... | ... | ... | ( | ) |
| 2 | 4½ | 7 | 9½ | 12 | ... | ... | ... | ... | ... | ... | ... | ... | ( | ) |
| 1 | 2 | 4 | 7 | 11 | ... | ... | ... | ... | ... | ... | ... | ... | ( | ) |
| 25 | 20 | 16 | 13 | 11 | ... | ... | ... | ... | ... | ... | ... | ... | ( | ) |
| 5½ | 9½ | 13½ | 17½ | 21½ | ... | ... | ... | ... | ... | ... | ... | ... | ( | ) |
| 0 | 3 | 8 | 15 | 24 | ... | ... | ... | ... | ... | ... | ... | ... | ( | ) |
| 3 | 6 | 12 | 15 | 30 | ... | ... | ... | ... | ... | ... | ... | ... | ( | ) |

## TEST 2c.—ANALOGIES.

This is like Test 1c on page 3. You may look back at the directions if you wish.

You UNDERLINE ONE word in each bracket.

coal : fire—food : ? ... ... ... (plate, stomach, knife, shop, pot)

square : circle—cube : ? ... ... (triangle, oblong, sphere, block, line)

wax : wane—stretch : ? ... ... (burst, swell, shape, shrink, round)

rail : wheel— ? : foot ... ... (leg, ground, spoke, hand, shoe)

bird : cage— ? : kennel ... ... (straw, chain, roof, dog, cat)

vegetable : carrot— ? : banana ... (yellow, skin, fruit, apple, red)

decision : hesitation—certainty : ? ... (thought, perplexity, assurance, judgment, relief)

contempt : admiration— ? : admire... (despise, condemn, ridicule, loathe, ignore)

## TEST 2d.—REASONING.

**Remember that you underline the right answer.**

A wire fence is supported by posts, spaced at a distance of one yard from each other. If there are twenty such posts on this fence, what is the distance in yards from the first post to the twentieth one? ... ... ... ... (nineteen, twenty, twenty-one)

Pit ponies are said to become blind through working underground. John has a pony which is blind. Was it formerly a pit pony? ... ... ... ... ... (Yes, No, I cannot tell)

In Willie's home there are his father and mother, his two sisters, and one brother. How many males are there in the household? ... ... ... ... ... ... (one, two, three)

How many daughters has Willie's father? ... ... ... (one, two, three)

Mr. Wilson is not healthy, and cannot travel for more than three hours at a stretch. He also feels sick if he is in a train for more than two hours, or in an aeroplane for more than one. If a motor car travels at 30 miles an hour, a train at 50 miles an hour, and an aeroplane at 80 miles an hour, which should Mr. Wilson use for a non-stop journey of 95 miles? ... ... ... ... (motor car, train, aeroplane)

In a foreign seaport, an Englishman who could speak only English wished to speak to a Chinaman, who knew only Chinese. The Englishman obtained a Frenchman who spoke both French and English, and a Russian who could speak both Russian and French. Meanwhile, the Chinaman met a fellow Chinaman who could speak both Russian and Chinese. Would the Englishman now be able to converse with the Chinaman?... ... ... (Yes, No, I cannot tell.

**Go on to NEXT PAGE without waiting to be told.**

## TEST 2e.—LOGICAL SELECTION.

Underline the TWO words in the bracket which tell what the thing outside the bracket is certain, or most likely, to have or to be connected with. This is like Test 1e on Page 4.

ship ... (sails, engine, hull, funnel, rudder)

enthusiasm (energy, patience, thought, zeal, nobility)

respect ... (malice, hatred, obedience, love, envy)

field ... (grass, earth, hedge, area, trees)

journey ... (return, departure, route, result, object)

## TEST 3a.—FOLLOWING DIRECTIONS.

**Remember to write the answer to the question in the bracket.**

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

If the alphabet were written backwards, write the letter which would be 8th in the new order ... ... ... ... ... ... ... ... ... ... ... ( )

If, after the alphabet written backwards, the alphabet was written again in the correct way, which letter would have most letters between its two appearances? ... ( )

Write the letter which would have six letters between its two appearances ... ... ( )

If there are more I's in DIMINISHING than in TRINITARIAN write P, unless there are more N's in the second than in the first, in which case write R ... ... ( )

If the letters of the word TROPICAL occur in the order opposite to their order in the alphabet, write O, but if not, put them in that order, and write here the letter which has to be shifted farthest ... ... ... ... ... ... ... ( )

If the 2nd, 4th, 6th and all the other even letters of the alphabet were lost, write what would remain of the word RINTINTIN ... ... ... ... ( )

## TEST 3b.—NUMBER SERIES.

**Remember to write in the bracket the number that should come next.**

624 312 156 78 ... ... ... ... ... ... ... ... ... ... ( )

19 18 20 19 21 ... ... ... ... ... ... ... ... ... ( )

4 8 16 32 ... ... ... ... ... ... ... ... ... ... ( )

216 36 6 1 ... ... ... ... ... ... ... ... ... ... ( )

1 3 6 10 15 ... ... ... ... ... ... ... ... ... ( )

63 42 25 12 ... ... ... ... ... ... ... ... ... ... ( )

1 $\frac{3}{2}$ $\frac{9}{4}$ $\frac{27}{8}$ $\frac{81}{16}$ ... ... ... ... ... ... ... ... ( )

1 4 8 11 22 ... ... ... ... ... ... ... ... ... ( )

**Go on to NEXT PAGE without waiting to be told.**

## TEST 3c.—ANALOGIES.

**Remember to underline the right word in each bracket.**

sacred : secular— ? : hall ... ... (mansion, church, window, cemetery, door)

uncle : nephew—aunt : ? ... ... (uncle, daughter, cousin, niece, son)

receipt : bill— ? : debt ... ... (account, credit, transaction, payment, money)

advance : retire—ascend : ? ... ... (climb, hill, withdraw, descend, retreat)

evolution : revolution—fire : ?... ... (smoke, burning, explosion, heat, matches)

migration : birds— ? : words ... (syllables, alteration, dictionary, translation, sentence)

## TEST 3d.—REASONING.

**Remember that you underline the right answer.**

Mr. Jones' watch always stopped when the hands passed each other, and then he had to restart it. If it was going at 6 a.m., how many times would he have to restart it by 6 p.m. ... ... ... ... ... ... ... (eleven, twelve, thirteen)

Mr. Brown is middle aged and can walk three miles in an hour. His son is young and can walk four miles in an hour, while his father, the son's grandfather, is so old that he can walk only two miles in an hour. The three go out together for an hour's walk. How far will they go? (two miles, three miles, four miles)

A Wolf Cub gains a star for each year he has been a Cub. Johnny who has been a Cub for nearly four years found a jersey of his which had one star on it. What is the least number of years he can have had this jersey? ... (one, two, three)

Mr. Brown came home from holiday and left the key of the home letter box at the seaside with his wife. He asked her to post it to him, and she did. Was that sensible, or silly ... ... ... ... ... ... ... (sensible, silly, I cannot tell)

John, Tom and Dick all like reading. John likes detective stories and school stories; Tom likes school stories and exploration stories; and Dick likes detective stories and exploration stories. The cleverest of these boys does not read many detective stories. Which is he? (John, Tom, Dick)

**THE END. Look over your work again.**

The frequencies of correct answers and the score-scatter were extracted by the author from the papers of 202 candidates in an English borough. In the course of the work the answer-pattern obtained from 97 candidates was found and compared with that obtained from other 105 candidates. The results of this comparison were shown on pages 51-59, and the separate answer-patterns graphed on pages 52. The answer-pattern for the complete group of 202 candidates is graphed on page 167, and is followed by the histogram of the score-scatter.

The skewness of the answer-pattern-differential was +0.27 and that of the score-scatter was -0.06. The value of h was 0.11 .

Table 41. Frequencies of correct answers to M.H.T. 9.
( 202 candidates )

| | Subtests | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | 1a | 1b | 1c | 1d | 1e | 2a | 2b | 2c | 2d | 2e | 3a | 3b | 3c | 3d |
| 1 | 192 | 176 | 119 | 155 | 142 | 177 | 79 | 82 | 72 | 16 | 174 | 71 | 23 | 29 |
| 2 | 188 | 146 | 96 | 160 | 112 | 47 | 98 | 47 | 97 | 13 | 16 | 66 | 99 | 81 |
| 3 | 189 | 146 | 106 | 167 | 118 | 27 | 64 | 46 | 127 | 53 | 5 | 102 | 62 | 13 |
| 4 | 111 | 130 | 56 | 168 | 92 | 163 | 68 | 57 | 161 | 39 | 117 | 14 | 68 | 115 |
| 5 | 180 | 141 | 66 | 139 | 93 | 84 | 11 | 144 | 20 | 47 | 16 | 72 | 19 | 120 |
| 6 | 75 | 62 | 81 | 179 | 7 | 103 | 89 | 125 | 56 | - | 23 | 6 | 8 | - |
| 7 | 150 | - | 90 | - | - | 137 | 50 | 40 | - | - | - | 37 | - | - |
| 8 | 175 | - | 80 | - | - | 129 | 36 | 57 | - | - | - | 14 | - | - |

Table 42. Score-scatter of M.H.T. 9. ( 202 candidates)

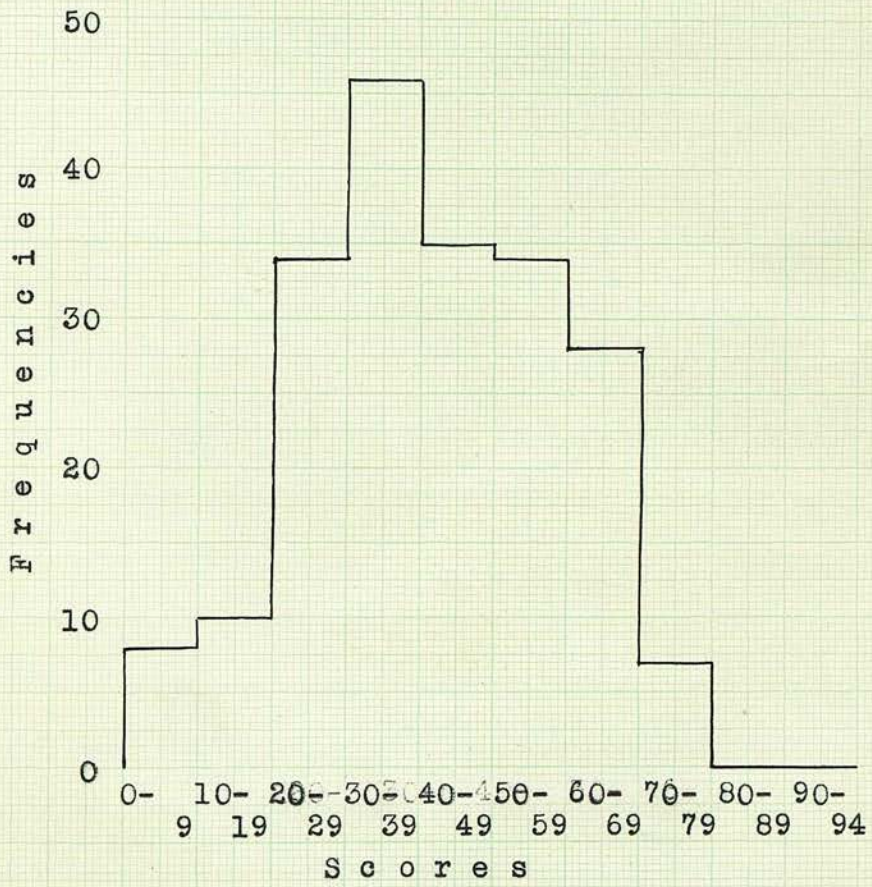
| Score | Frequency | Score | Frequency | Score | Frequency |
|---|---|---|---|---|---|
| 0 | - | 26 | 3 | 52 | 3 |
| 1 | - | 27 | 2 | 53 | 4 |
| 2 | 2 | 28 | 7 | 54 | 5 |
| 3 | 1 | 29 | 2 | 55 | 1 |
| 4 | 1 | 30 | 4 | 56 | 6 |
| 5 | 2 | 31 | 4 | 57 | 1 |
| 6 | - | 32 | 4 | 58 | 4 |
| 7 | - | 33 | 7 | 59 | 1 |
| 8 | 2 | 34 | 4 | 60 | 2 |
| 9 | - | 35 | 5 | 61 | 2 |
| 10 | - | 36 | 5 | 62 | 3 |
| 11 | - | 37 | 5 | 63 | 3 |
| 12 | - | 38 | 4 | 64 | 3 |
| 13 | 2 | 39 | 4 | 65 | 1 |
| 14 | - | 40 | 5 | 66 | 3 |
| 15 | 1 | 41 | 10 | 67 | 4 |
| 16 | 1 | 42 | 4 | 68 | 3 |
| 17 | 1 | 43 | 5 | 69 | 4 |
| 18 | 1 | 44 | 2 | 70 | 2 |
| 19 | 4 | 45 | 1 | 71 | 2 |
| 20 | 3 | 46 | - | 72 | 1 |
| 21 | - | 47 | - | 73 | 1 |
| 22 | 4 | 48 | 5 | 74 | 1 |
| 23 | 3 | 49 | 3 | 75 and over | - |
| 24 | 7 | 50 | 2 | | |
| 25 | 3 | 51 | 7 | | |

In the more usual form of grouped scores the table is

| Score | |
|---|---|
| 0 - 9 | 8 |
| 10 - 19 | 10 |
| 20 - 29 | 34 |
| 30 - 39 | 46 |
| 40 - 49 | 35 |
| 50 - 59 | 34 |
| 60 - 69 | 28 |
| 70 - 79 | 7 |
| 80 - 89 | 0 |
| 90 - 94 | 0 |
| | 202 |

Figure 28. Answer-pattern of M.H.T. 9. ( 202 candidates

Figure 29. Score-scatter of M.H.T. 9 (202 candidates)



M. H. T. 11.

| Not to be filled in by the Scholar. | |
|---|---|
| Age on 1/8/31. years. | months. |
| Raw Score | |
| I.Q. | |

**DO NOT OPEN THIS BOOK UNTIL YOU ARE TOLD.**

EDUCATION COMMITTEE.

JUNIOR SCHOLARSHIP AND SECONDARY SCHOOL ENTRANCE EXAMINATION.

# INTELLIGENCE TEST.

10.5 to 10.15 a.m.—DISTRIBUTION, PRELIMINARY INSTRUCTIONS AND ENTRIES.

10.15 to 11.0 a.m.—WORKING OF TEST.

**Fill in the following particulars at once:—**

Your full name.....

Name of the School you attend.....

Standard, Class or Form you are in.....

**Read the following carefully:—**

1. When you are told to begin, answer the questions as quickly and as carefully as you can.
2. Begin at the beginning and go straight through.
3. If you cannot do any question in any test, leave it out and go on to the next.
4. When you finish one page, go on to the next.
5. You will have 45 minutes. No one is expected to do everything. Just do as much as you can.
6. **Ask no questions at all.**

**Read each question carefully and then answer it in the bracket. Begin at the beginning and go straight through. Try each question as you come to it ; but if you cannot do it soon, go on to the next. The alphabet is printed here to help you with some of the questions.**

---

**A B C D E F G H I J K L M N O P Q R S T U V W X Y Z**

---

1. Do you understand that you must do your best and not ask questions ? If so write W in the bracket ... ... ... ... ... ... ... ... ... ( )
2. Write the letter which comes before R in the alphabet ... ... ... ... ( )
3. Write the numbers between 2 and 8 and then cross out the middle one ( )
4. If $\frac{1}{7}$ is larger than $\frac{1}{8}$ write P, if not write C ... ... ... ... ... ... ( )
5. If 25d. is the same as 2/1 write F, if not write M ... ... ... ... ... ( )

Look at this example :—

Finger is to hand as toe is to what ? ... ... ... ... (foot, knee, arm, shoe, nail).

The answer is one of the words in the bracket. The right answer is foot, and it is underlined.

Now look at this next example, and then underline the right answers in questions 6, 7, and 8.

Example :—Man is to clothes as what is to fur ? ... ... (coat, animal, bird, skin, cloth).

6. Apple is to fruit as carrot is to ... ... ... (soup, dessert, vegetable, dish, garden).
7. School is to teacher as church is to ... ... (student, minister, pupil, old, spire).
8. Soldier is to army as what is to navy ? ... ... (captain, singer, gun, sailor, trench).
9. Write the letter which comes most often in the word Tennessee ... ... ( )
10. If T and S come together in the alphabet write K, if not write D ... ... ( )
11. Fill in the number which has been rubbed out in the top line of this multiplication sum, and write it in the bracket as well.

$$\begin{array}{r} 2\,.\,3 \\ \times\quad 5 \\ \hline 1\,2\,1\,5 \\ \hline \end{array}$$

... ... ... ... ... ... ( )

12. Water is to boat as railway is to ... (sea, aeroplane, submarine, automobile, train)
13. John is younger than Jim, and Jim is younger than Bill. Which is the youngest of the three ? ... ... ... ... ... ... ... (John, Jim, Bill).

    (Do not write anything, just underline the right one in the bracket).

14. Coke makes better fuel than wood, but not so good as coal. Which makes the best fuel ? ... ... ... ... ... ... ... (Coke, wood, coal).
15. If E and F changed places in the alphabet, what would the 6th letter be ? ... ( )
16. If E is found before H in the alphabet and S is found before M, write T. But if only one of these is true write O ... ... ... ... ... ... ... ( )

**Go on to NEXT PAGE without waiting to be told.**

17. Look at these three proverbs. Two of them mean nearly the same. Put a cross plainly after the other one.

Distance lends enchantment to the view.
Too many cooks spoil the broth.
Absence makes the heart grow fonder.

18. Do the same with these three. Find which of them mean nearly the same, and then put a cross after the other one.

He who hesitates is lost.
Faint heart ne'er won fair lady.
Look before you leap.

19. Fill in the missing number in this subtraction sum, and write it in the bracket as well.

2 . 7 3 4 ... ... ... ... ... ... ( )
1 4 8 2 5
———————
1 3 9 0 9
———————

20. Chalk is to blackboard as pencil is to ... ... ... (grass, ink, paper, dust, point)

21. I have three children, Mary, Jim, and Mabel. Mary is taller than Jim, but Mabel is tallest of all. Which is the shortest ? ... ... (Mary, Jim, Mabel)

Look at these five words:—Dog, elephant, sparrow, cow, lion. One of them is different from the other four, and so it is underlined. Sparrow is different because all the others are animals.

Look at this example:—Hot, freezing, warm, cold, wet.

Here wet is different because all the others are about temperature. Now underline the "different" word in these five:—

22. Chair, book, couch, bed, bench.

Do the next three in the same way:—

23. Dog, cheese, potato, bread, marmalade.

24. Iron, silver, wool, copper, zinc.

25. Banana, plum, apple, orange, table.

26. Feathers are to bird as fur is to ... ... ... ... ... (cap, chicken, cat, tree, egg)

27. Day is to night as what is to darkness ? ... ... (early, light, noon, sun, star).

28. Write the letter which is midway in the word BEFRIEND between the two letters which are the same ... ... ... ... ... ... ... ... ... ( )

29. MURDNUNOC is a word written backward. Write it as it usually appears.

( )

30. If the letter G occurs most often in the word GIGGLING write the middle letter of the word PIP, unless O and N come next to one another in the alphabet, in which case write W instead ... ... ... ... ... ... ... ... ( )

31. Sting is to bee as what is to soldier ? ... ... (belt, bayonet, pen, uniform, helmet)

**Go on to NEXT PAGE without waiting to be told.**

Underline the " different " word in each of the next three questions :—

32. Sea, river, lake, rock, pool.

33. Hat, room, ribbon, dress, coat.

34. Horse, squirrel, zebra, fence, pig.

In a certain secret writing

| FNRH | TNPRBUTZNDNRLF | WKPZQEC | means |
|---|---|---|---|
| s e n d | r e i n f o r c e m e n t s | q u i c k l y | |

35. In the same secret writing you find this. Write below it what it means :—
ZUDN FUUR LU EURN LTNN

36. Wheat is dearer than barley, and barley is cheaper than oats. Which is the cheapest ?
( )

37. John has two sisters, Annie and Edna, and two brothers, Harold and Wade, and three cousins, Dick, Bill, and their sister Florence. How many male cousins has Bill ? ... ... ... ... ... ... ... ... ... ... ... ( )

38. How many female cousins has Annie ? ... ... ... ... ... ... ( )

39. Hear is to telephone as see is to ... (shout, telegraph, spyglass, distance, receiver)

40. Suppose every fifth letter of the alphabet (E, J, and so on) were lost, what would then be the tenth letter ? ... ... ... ... ... ... ... ... ( )

Underline the " different " word in each of the next three questions :—

41. Run, jump, skate, sit, slide.

42. Mast, deck, funnel, quay, rudder.

43. Pistol, spear, helmet, sword, rifle.

44. Fill in the missing numbers in this multiplication sum, and write them in the brackets also, opposite the proper lines :—

```
   2 . 3   ... ... ... ... ... ... ( )
     4 6
  -------
   1 6 . 8 ... ... ... ... ... ... ( )
 1 . 9 2   ... ... ... ... ... ... ( )
  -------
 1 2 5 5 8
  -------
```

45. Do you understood that you have to go on trying till time is up ? If so, write T ( )

46. If I am facing the north with my arms stretched sideways, in what direction is my right arm pointing ? ... ... ... ... ... ... ... ... ( )

**Go on to NEXT PAGE without waiting to be told.**

47. Two of these proverbs have somewhat similar meanings. Mark the other one with a cross :—

Pride goes before a fall .
A stitch in time saves nine.
Prevention is better than cure.

48. Do the same with these :—

Honesty is the best policy.
Those who live in glass houses shouldn't throw stones.
The pot shouldn't call the kettle black.

49. Underline the word in the bracket which means nearly the same as crooked ... ... ... ... (straight, large, round, bent, wide)

Do the same with

50. boast ... ... ... ... ... ... ... (brag, speak, explain, tell, relate)

51. brief ... ... ... ... ... ... ... (resounding, tall, heavy, short, black)

Cross out, with an X, plainly, the word in the bracket which is nearly the opposite of

52. agree ... ... ... ... ... ... ... (travel, quarrel, love, like, meet)

53. rough ... ... ... ... ... ... (hilly, handsome, smooth, cheat, bargain)

54. The words in this sentence have been mixed up. Write it out as it ought to be, beneath the printed sentence :—

EARTH IS MINED COAL THE FROM

55. Do the same with this :—

PROPERTY FLOODS LIFE AND DESTROY.

56. Fill in the missing numbers in this division sum, and also write the numbers in the brackets opposite their lines :—

```
43 ) 2 8 9 . 1 ( 6 7 3   ... ... ... ... (  )
     2 5 8
     -----
       3 1 . ...  ... ... ... ... ... (  )
       3 0 . ...  ... ... ... ... ... (  )
       -----
         1 . 1    ... ... ... ... ... (  )
         1 . 9    ... ... ... ... ... (  )
         -----
             2
         -----
```

57. Write the two letters in the word BRUSH which have as many letters between them in the alphabet as there are letters in the word BRUSH itself ... (  )

58. Three posts are in an exact straight line, and from where I am standing I can only see one of them because the others are exactly behind it. I now move six steps to the left, so that I can see them all. Which is the farthest away, the right-hand one, the middle one, or the left-hand one ? ... ... ... ... ... (right, middle, left)

59. One month of the year begins with the letter which comes before T in the alphabet.
Write the first letter of the month before this one ... ... ... ... ... ( )

Look at the word in front of the bracket, and in the bracket find a word which is either nearly the same, or nearly the opposite. Underline it if it is the same, cross it out with an X if it is opposite.

60. summit ... ... ... ... ... ... (ascend, nothing, top, compass, add)

61. chase ... ... ... ... ... ... ... (silver, please, find, pursue, take)

62. yes ... ... ... ... ... ... ... ... (please, thanks, what, no, oh)

63. often ... ... ... ... ... ... (before, seldom, behind, aside, soon)

64. perpetual ... ... ... ... ... (endless, partial, inevitable, frequent, fluent)

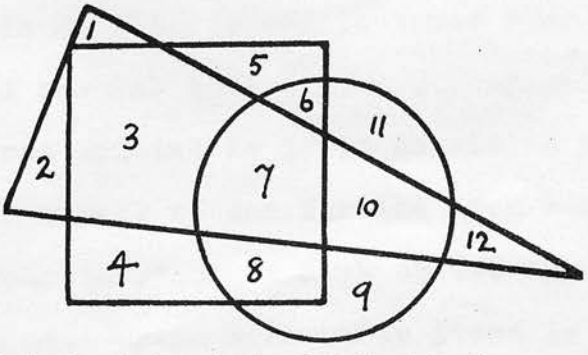
65. transparent ... ... ... (occasional, inheritance, granular, polished, opaque

The next question is written in the secret writing you have already seen in question 35. Find out what it means and answer it. You can get most of the letters from the explanation in front of question 35, but I must also tell you that G in the secret writing means "a" in ordinary writing; and there are two letters you will have to guess.

66. GTN CUK G IUC UT G YPTE? ... ... ... ... ( )

67. Write this sentence as it ought to read:—

AND EMOTIONS SORROW SIMILAR GRIEF ARE




68. Which number is in the circle and triangle but not in the square? ... ... ( )

69. Subtract the number which is in the square and triangle but not in the circle, from the sum of all the numbers which are in the circle but outside the triangle ... ... ... ... ... ... ... ... ... ... ... ( )

70. Add together all the numbers each of which is in two only of the figures ... ( )

71. It is eleven minutes to nine. What time would you think it was if you mistook the long hand of the clock for the short hand, and the short hand for the long?

**THE END. Look over your work again.**

The data given below for M.H.T. 11 were extracted from the papers of 209 candidates in an English borough. At the time about 1500 papers were available, but it was found that the answer-pattern obtained from the first batch of 104 papers agreed very well with that obtained from the second set of 105 papers, and therefore only 209 papers were analysed. The comparison of the separate answer-patterns is described in Part I, page 48, and the answer-patterns are graphed on page 50. The answer-pattern for the combined group is graphed after the tables below, as is also the score-scatter.

The skewness of the answer-pattern-differential was +0.15 and that of the score-scatter was -0.15; h was .10 .

The calculation of the answer-pattern was made a little more difficult than usual by the use in this test of items carrying multiple marks. The items concerned were numbers 28, 35, 47, 48, 51, 64, 66, 67, 69, 70, 71. The method used to surmount this difficulty was to treat such an item as the aggregate of several unit items. For example an item bearing four marks was treated as if it consisted of four separate items. The award of a mark of one for the item was credited to the first of these separate items, a mark of two was credited to the first two, and so on. There were other items in the test which appear in the table below as if they bore multiple marks, but in their case they did consist of separate, though not always independent items. This interlinking of items would reduce the amount of

hig in the test.

Another interesting point in the results was the sex difference shown in the responses to question 31. Correct replies were given by 27 boys out of 48 in the first group, and by 45 out of 69 in the second, whereas the number of girls answering correctly was 9 out of 57 in the first group and 9 out of 35 in the second. With such pitfalls is the path of the examiner strewn !

Table 43. Frequencies of correct answers to M.H.T. 11.

| Question | Frequency | | |
|---|---|---|---|
| 1 | 203 | 26 | 147 |
| 2 | 203 | 27 | 146 |
| 3 | 148 | 28 | 95 |
| 4 | 145 | | 94 |
| 5 | 191 | 29 | 155 |
| 6 | 193 | 30 | 136 |
| 7 | 189 | 31 | 90 |
| 8 | 175 | 32 | 179 |
| 9 | 176 | 33 | 181 |
| 10 | 184 | 34 | 177 |
| 11 | 102 | 35 | 96 |
| 12 | 203 | | 86 |
| 13 | 161 | 36 | 182 |
| 14 | 179 | 37 | 75 |
| 15 | 179 | 38 | 124 |
| 16 | 188 | 39 | 134 |
| 17 | 98 | 40 | 103 |
| 18 | 78 | 41 | 181 |
| 19 | 139 | 42 | 143 |
| 20 | 184 | 43 | 146 |
| 21 | 201 | 44 | 74 |
| 22 | 169 | | 122 |
| 23 | 180 | | 62 |
| 24 | 173 | 45 | 202 |
| 25 | 184 | 46 | 111 |

| | | | |
|---|---|---|---|
| 47 | 47 | 63 | 90 |
| | 42 | 64 | 40 |
| 48 | 58 | | 36 |
| | 55 | 65 | 46 |
| 49 | 181 | 66 | 60 |
| 50 | 144 | | 60 |
| 51 | 63 | | 60 |
| | 63 | | 14 |
| 52 | 116 | 67 | 43 |
| 53 | 107 | | 43 |
| 54 | 75 | | 43 |
| 55 | 117 | | 43 |
| 56 | 29 | 68 | 113 |
| | 27 | 69 | 25 |
| | 69 | | 25 |
| | 29 | | 25 |
| | 38 | | 25 |
| 57 | 77 | 70 | 9 |
| 58 | 74 | | 9 |
| 59 | 79 | | 9 |
| 60 | 73 | | 9 |
| 61 | 92 | 71 | 32 |
| 62 | 98 | | 32 |

Table 44. Score-scatter of M.H.T. 11.

| Score | Frequency | | | | |
|---|---|---|---|---|---|
| 0 | 1 | 17 | 2 | 34 | 2 |
| 1 | - | 18 | 2 | 35 | 4 |
| 2 | 1 | 19 | 1 | 36 | 1 |
| 3 | - | 20 | 4 | 37 | 5 |
| 4 | - | 21 | - | 38 | 3 |
| 5 | - | 22 | - | 39 | 5 |
| 6 | - | 23 | 2 | 40 | 3 |
| 7 | 1 | 24 | 1 | 41 | 6 |
| 8 | 1 | 25 | 2 | 42 | 5 |
| 9 | - | 26 | 2 | 43 | 1 |
| 10 | - | 27 | 2 | 44 | 3 |
| 11 | 1 | 28 | 2 | 45 | 4 |
| 12 | - | 29 | 3 | 46 | 4 |
| 13 | - | 30 | 5 | 47 | 8 |
| 14 | 1 | 31 | 1 | 48 | 3 |
| 15 | - | 32 | 5 | 49 | 5 |
| 16 | - | 33 | 4 | 50 | 2 |

| | | | | | |
|---|---|---|---|---|---|
| 51 | 7 | 66 | 1 | 81 | - |
| 52 | 3 | 67 | 4 | 82 | 1 |
| 53 | 9 | 68 | 1 | 83 | 1 |
| 54 | 4 | 69 | 2 | 84 | 1 |
| 55 | 3 | 70 | 4 | 85 | - |
| 56 | 3 | 71 | 3 | 86 | - |
| 57 | 4 | 72 | 5 | 87 | - |
| 58 | 3 | 73 | 1 | 88 | 1 |
| 59 | 5 | 74 | 3 | 89-96 | - |
| 60 | 3 | 75 | 3 | | |
| 61 | 4 | 76 | 1 | | 209 |
| 62 | 3 | 77 | 1 | | |
| 63 | 4 | 78 | 1 | | |
| 64 | 5 | 79 | 1 | | |
| 65 | 3 | 80 | 3 | | |

Grouped in intervals of ten marks, the score-scatter is

| Score | Frequency |
|---|---|
| 0 - 9 | 4 |
| 10 - 19 | 7 |
| 20 - 29 | 18 |
| 30 - 39 | 35 |
| 40 - 49 | 42 |
| 50 - 59 | 43 |
| 60 - 69 | 30 |
| 70 - 79 | 23 |
| 80 - 89 | 0 |
| 90 - 96 | 0 |
| | 209 |

Figure 30. Answer-pattern of M.H.T. 11.
( 209 candidates )

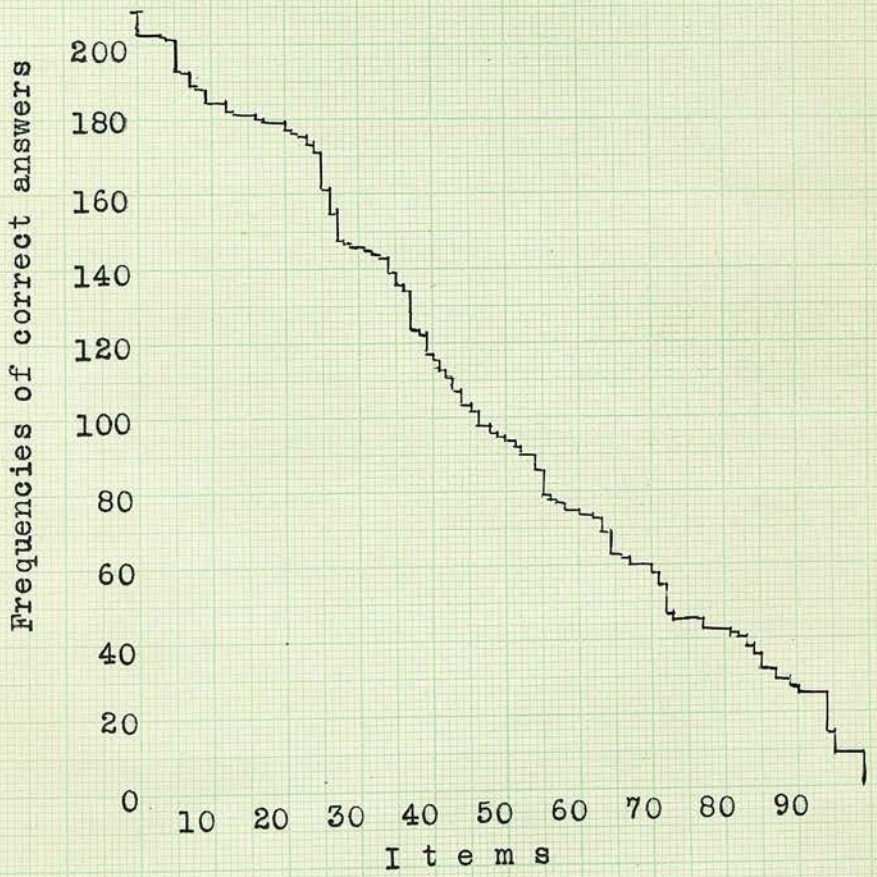
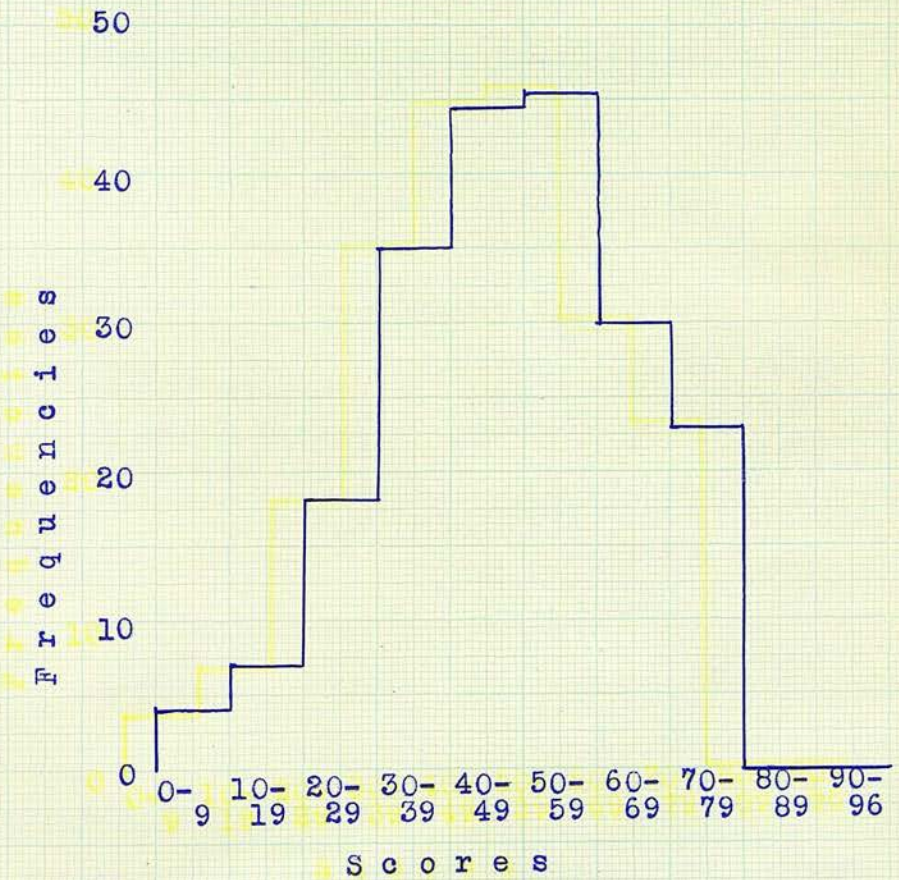


Figure 31. Score-scatter of M.H.T. 11. (209 candidates)



M. H. T. 12.

( Mental Survey Form )

DO NOT OPEN THE BOOK UNTIL YOU ARE TOLD.

# THE SCOTTISH COUNCIL FOR RESEARCH IN EDUCATION

SEX
(indicate by X).

| Boy. |
|---|
| |
| Girl. |
| |

## MENTAL SURVEY

### INSTRUCTIONS TO PUPILS

Listen carefully to the teacher and do quickly and carefully what you are told to do.

Surname: Christian Names:

Name of Pupil in block capitals, Surname first } .....

| Name of County. | Burgh or Parish. | School. |
|---|---|---|
| | | |

Date of Birth.*

| Day. | Month. | Year. |
|---|---|---|
| | | |

**To be completed by Teacher.**

| Class in School. |
|---|
| Post Primary V..... |
| IV..... |
| III..... |
| II..... |
| I..... |
| Senior Upper..... |
| Middle..... |
| Lower..... |
| Junior Upper..... |
| Lower..... |
| Infants..... |
| Special ..... |

**FOR MARKER'S USE ONLY.**

PICTURE TESTS.

| Page. | Score. | Marked by. |
|---|---|---|
| 2 | | |
| 3 | | |
| Total, Picture Tests. | | |

Checked by .....

Entered in nominal roll by .....

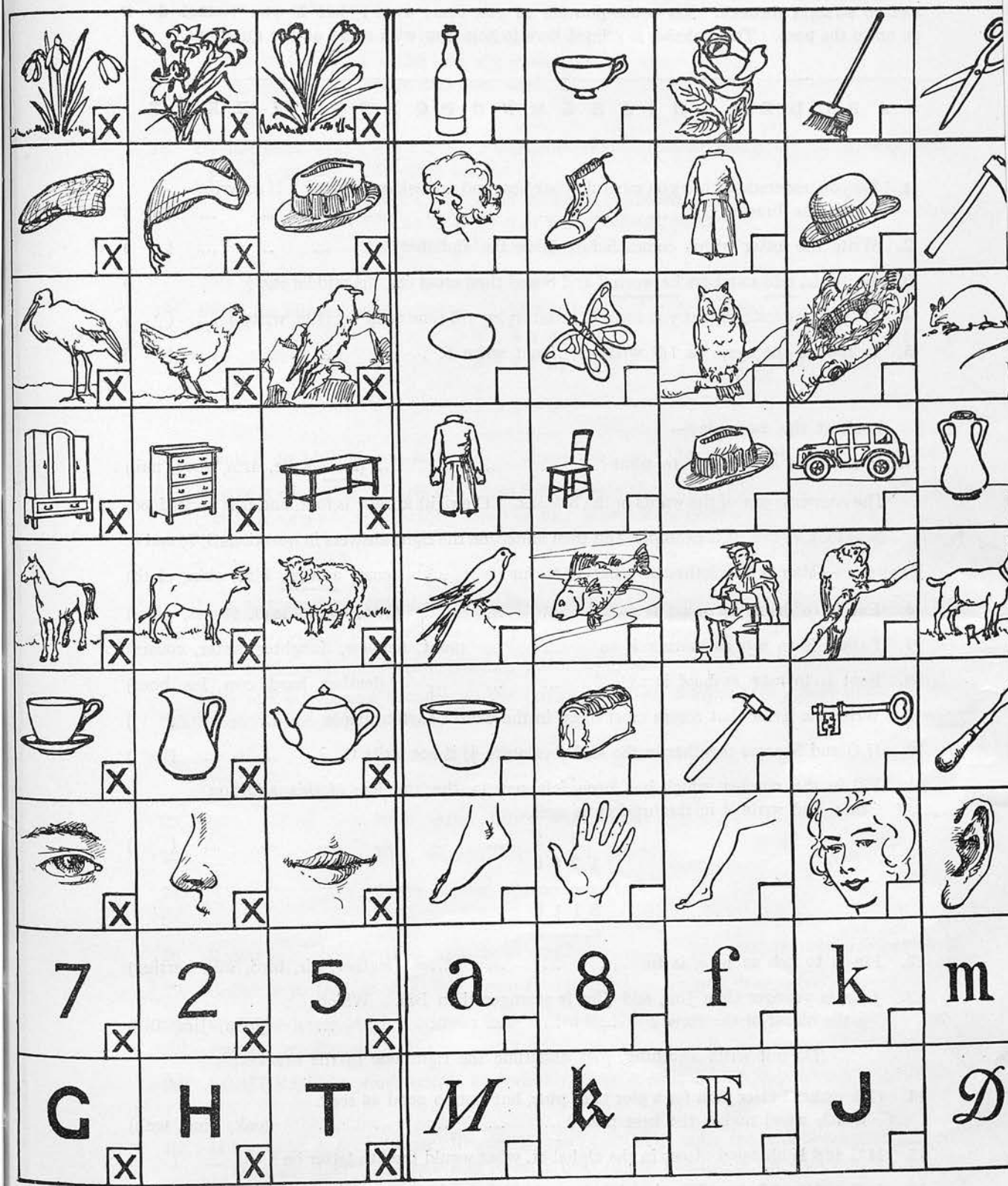
Tabulated by .....

VERBAL TESTS.

| Page. | Score. | Marked by | |
|---|---|---|---|
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| Total of pages 4 to 8. | | | |

* To be checked from register by Teacher.

1 2 3 4 5




SCORE.....

**Read each question carefully and then answer it in the bracket. Begin at the beginning and go straight through. Try each question as you come to it; but if you cannot do it go on to the next. The alphabet is printed here to help you with some of the questions.**

---

**A B C D E F G H I J K L M N O P Q R S T U V W X Y Z**

---

1. Do you understand that you must do your best and not ask questions? If so write L in the bracket ... ... ... ... ... ... ... ... ... ( )
2. Write the letter which comes before M in the alphabet ... ... ... ... ( )
3. Write the odd numbers between 2 and 8 and then cross out the middle one ( )
4. Do you understand that you have to go on trying till time is up? If so, write 5 ... ( )
5. If 19d. is the same as 1/7 write G, if not write R ... ... ... ... ( )

Look at this example:—

Finger is to toe as hand is to what? ... ... ... ... (foot, knee, arm, shoe, nail)

The answer is one of the words in the bracket. The right answer is foot, and it is underlined.

Now look at this next example, and then underline the right answers in questions 6, 7, and 8.

EXAMPLE:—Man is to clothes as what is to fur? ... (coat, animal, bird, skin, cloth)

6. Eat is to drink as bread is to ... ... ... (iron, water, lead, stones, grass)
7. Father is to son as mother is to ... ... (aunt, nephew, daughter, sister, cousin)
8. Foot is to man as hoof is to ... ... ... ... (leather, hard, cow, leg, boot)
9. Write the letter that comes most often in the word Constantinople ... ... ( )
10. If O and N come together in the alphabet write J, if not write C ... ... ... ( )
11. Fill in the number which has been left out in the top line of this addition sum, and write it in the bracket as well.

```
  3 . 2
1 7 6 6
-------
2 1 1 8
-------
```

... ... ... ... ... ... ( )

12. Fin is to fish as wing is to ... ... ... ... (feather, air, bird, sail, herring)
13. John is younger than Jim, and Jim is younger than Bill. Which is the oldest of the three? ... ... ... ... ... ... (John, Jim, Bill)

(Do not write anything, just underline the right one in the bracket).

14. Oak makes better piles for a pier than pine, but not so good as teak. Which wood makes the best piles? ... ... ... ... ... (oak, pine, teak)
15. If G and H changed places in the alphabet, what would the 8th letter be? ... ( )
16. If G is found before J in the alphabet and R is found before L, write S. But if only one of these is true write P ... ... ... ... ... ... ... ( )

**Go on to NEXT PAGE without waiting to be told.**

17. Look at these three proverbs. Two of them mean nearly the same. Put a cross plainly after the other one.

Well begun is half done.
It's the first step that counts.
Waste not want not.

18. Do the same with these three. Find which of them mean nearly the same, and then put a cross after the other one.

There's a skeleton in every cupboard.
It's an ill wind that blows nobody good.
Every cloud has a silver lining.

19. Fill in the missing number in this subtraction sum, and write it in the bracket as well.

3 . 8 4 5 ... ... ... ... ... ... ( )
2 5 9 3 6
---
1 3 9 0 9
---

20. Duck is to bird as iron is to ... ... ... ... (water, goose, metal, steel, lead)

21. I have three uncles. Uncle Fred lives farther away than Uncle Alec, but Uncle Jack lives farthest away of all. Which lives nearest to me? ... ... ... ... ... ... ... ... ... (Fred, Alec, Jack)

Look at these five words :—Dog, elephant, sparrow, cow, lion. One of them is different from the other four, and so it is underlined. Sparrow is different because all the others are animals.

Look at this example :—Hot, freezing, warm, cold, wet.

Here wet is different because all the others are about temperature. Now underline the "different" word in these five :—

22. Right, night, bright, black, fright.

Do the next three in the same way :—

23. Rain, water, calico, wine, milk.

24. Boy, waggon, kitten, girl, puppy.

25. Knife, saucer, spoon, fork, tart.

26. Meeow is to bow-wow as what is to dog? ... (hen, cat, donkey, speech, bark)

27. Bullet is to lead as what is to gold? ... ... (paper, coin, silver, copper, purse)

28. Write the letter which is midway in the word BLUEBIRD between the two letters which are the same ... ... ... ... ... ... ... ... ... ( )

29. EGAIRRAC is a word written backward. Write it as it usually appears—

( )

30. If the letter A occurs most often in the word CANADA write the middle letter of the word SLEEP unless P and R come next to one another in the alphabet, in which case write Y instead ... ... ... ... ... ... ... ... ( )

31. Dog is to terrier as what is to Liverpool? ... ... (city, cow, horse, state, cotton)

Underline the " different " word in each of the next three questions :—

32. Radiator, violin, flute, piano, saxophone.

33. Rain, snow, storm, mast, hail.

34. Sheep, lily, cart, trout, thrush.

In a certain secret writing

| | | | |
|---|---|---|---|
| lzqkcofu, | ftt r | yggr | means |
| STARVING, | NEED | FOOD | |

35. In the same secret writing you find this. Write below it what it means :—

yoct kgctkl rtqr.

36. " Tragu " is cheaper than " vashol," and " vashol " is dearer than " spongop." Which is the dearest? ... ... ... ( )

37. John's mother has no brothers or sisters. His father has a bachelor brother Frank, and a married sister Mary who has two daughters and one son (Annie, Elizabeth, and Timothy). How many aunts has John ? ... ... ... ... ... ( )

38. How many nieces has John's father ? ... ... ... ... ... ... ( )

39. Establish is to abolish as begin is to ... (work, year, end, commence, despair)

40. Suppose every fourth letter (D, H, L, and so on) were lost, what would then be the tenth letter ? ... ... ... ... ... ... ... ... ... ( )

Underline the " different " word in each of the next three questions :—

41. Sixpence, shilling, penny, farthing, franc.

42. Eye, pen, nose, chin, ear.

43. Cheap, sweet, sour, salty, bitter.

44. Underline the ONE of the four answers to each statement which seems to you to be correct :—

If your clothes catch fire—(roll yourself in rug or blanket, run about, 'phone fire-brigade, pour on petrol).

A window to ventilate properly must be—(made of stained glass, open top and bottom, polished with chamois leather, covered with curtains).

To prevent tools from rusting rub with—(sandpaper, tar, vaseline, file).

45. If $\frac{1}{6}$ is larger than $\frac{1}{5}$ write Q, if not write E ... ... ... ... ... ( )

46. If I am facing the west with my arms stretched sideways, in what direction is my left arm pointing ? ... ... ... ... ... ... ... ... ( )

47. Two of these proverbs have somewhat similar meanings. Mark the other one with a cross :–

Two heads are better than one.
Too many cooks spoil the broth.
Many hands make light work.

48. Do the same with these three. Find which of them mean nearly the same, and then put a cross after the other one.

Time and tide wait for no man.
It's an ill wind that blows nobody good.
Make hay while the sun shines.

49. Underline the word in the bracket which means nearly the SAME as little ... ... ... ... ... (large, round, small, bent, wide)

Do the same with :—

50. accept ... ... ... ... ... ... ... (take, give, hear, learn, find)

51. appeal ... ... ... ... ... ... (split, cleave, remind, beseech, revoke)

Cross out with an X, plainly, the word in the bracket which is nearly the OPPOSITE of

52. good .. ... ... ... ... ... ... (fine, bad, nice, clever, dark)

53. cautious ... ... ... ... ... ... (publish, appoint, suit, careful, heedless)

54. The words in this sentence have been mixed up. Write it as it ought to be, beneath the printed sentence :—

HUMP CAMEL HAS A HIS A BACK ON

55. Do the same with this :—

TRUE BOUGHT CANNOT FRIENDSHIP BE

56. Underline the ONE of the four answers to each statement which seems to you to be correct :—

Vitamine is found in—(fresh milk and fruits, lard, dried fruits, stale bread).

Metals can be joined together by—(gluing, riveting, nailing, polishing).

The forecastle of a ship is at the—(bow, stern, bridge, quarterdeck).

57. Write the two letters in the word TRENCH which have three letters between them in the alphabet ... ... ... ... ... ... ... ... ... ( )

58. Three posts are at the corners of an equilateral, that is an equal-sided, triangle. From where I am standing, the post nearest to me seems to be exactly half-way between the other two. If I now take two sidesteps to the left, will the posts look like this I I I

or like this? I I I

Mark the right one with tick ✓.

59. Which is the first month after Midsummer Day which has an **r** in its name? ... ... ... ... ... ... ( )

Look at the word in front of the bracket, and in the bracket find one word which is either nearly the same, or nearly the opposite. Underline it if it is the same, cross it out with an **X** if it is opposite.

60. jumps ... ... ... ... ... ... ... (leaps, runs, flies, swims, rests)

61. bring ... ... ... ... ... ... ... (take, think, make, mend, drop)

62. no ... ... ... ... ... ... ... (thanks, please, yes, perhaps, what)

63. fragrant ... ... ... ... ... (transparent, odorous, critical, brave, fragile)

64. legislature ... ... ... ... ... (executive, municipal, parliament, court, palace)

65. oscillate ... ... ... ... ... ... ... (bring, swing, king, sing, bright)

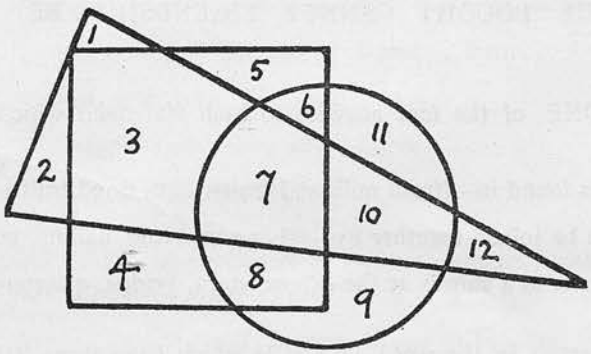
The next question is written in the secret writing you have already seen in question 35. Write down what it means and answer it. You can get most of the letters from the explanation in front of question 35, but there are some letters you will have to guess.

66. o l z g r q n D g f r q n ? ... ( )

Answer ... ... ... ( )

67. Write this sentence as it ought to read :—

BELL MOST TELEPHONES HAVE ATTACHED A




68. What number is in the triangle and square but not in the circle? ... ... ... ( )

69. What is the sum of the two numbers which are in the circle only?... ... ... ( )

70. Subtract the number which is in the circle and triangle but not in the square from the sum of all the numbers which are in the square but outside the circle ... ( )

71. If there are more I's in DIMINISHING than in TRINITARIAN write **P**, unless there are more N's in the second than in the first, in which case write **R**. ... ( )

**THE END. Look over your work on pages 4 to 8 again.**

The test M.H.T. 12 is particularly interesting as being that employed by the Scottish Council for Research in Education in their mental survey of Scotland in 1932. The test was applied in June 1932 to all the available children in Scotland b born in 1921. As it had been designed originally to test entrants for scholarships, the verbal test referred to as 12v was supplemented by a pictorial test so that the poorer candidates might have sufficient scope.

In an earlier form the picture portion of the test consisted of three parts; (1) a Picture-Digit Substitution Test (2) a Pictorial Classification Test, and (3) a Cube Counting Test. For reasons given in the report of the survey, "The Intelligence of Scottish Children", the Counting Cubes Test was dropped, and later, after the test had been given, the results of the Picture-Digit Substitution Test were ignored. The sole remaining picture test was thus the Pictorial Classification Test. It is the test referred to as M.H.T. 12p in the preceding pages.

The results now given have been published in the report mentioned above, but they are reproduced here for the sake of completeness. The papers from which were obtained the frequencies of correct answers were those of 500 boys and 500 girls in the verbal test, and of 450 boys and 450 girls in the picture test. These cases were selected "at random" by the marking Committees. The randomness of the selection may be tested to a certain extent

by comparing the average marks in the sample with the median marks for all Scotland. This is done in the following table reproduced from page 87 of "The Intelligence of Scottish Children".

| | Boys | | Girls | |
|---|---|---|---|---|
| Test | This Sample | Scotland | This Sample | Scotland |
| Verbal | 37.09 | 34.66 | 35.54 | 34.41 |
| Picture | 6.83 | 6.80 | 6.72 | 6.80 |

The skewness of the answer-pattern-differential for the verbal part, 12v, was found to be +0.199 , and that of the score-scatter was -0.17. The score-scatter was obtained from the whole population tested- some 87,000 boys and girls. The coefficient of hig was 0.16 . In the pictorial test 12p the skewness of the answer-pattern-differential was -1.201 and that of the score-scatter was -1.33 . The coefficient of hig was still fairly low, being 0.20. Here again the strong negative skewness of the answer-patterned differential is accompanied by a strong negative skewness of the score-scatter.

( Table )

Table 45. Frequencies of correct answers to M.H.T. 12.

Verbal Test

| Question | Number Correct ( out of 500) | |
|---|---|---|
| | Boys | Girls |
| 1 | 467 | 450 |
| 2 | 450 | 456 |
| 3 | 248 | 239 |
| 4 | 456 | 458 |
| 5 | 388 | 401 |
| 6 | 393 | 383 |
| 7 | 357 | 389 |
| 8 | 359 | 338 |
| 9 | 248 | 232 |
| 10 | 280 | 297 |
| 11 | 408 | 408 |
| 12 | 403 | 391 |
| 13 | 396 | 386 |
| 14 | 316 | 315 |
| 15 | 322 | 300 |
| 16 | 382 | 393 |
| 17 | 168 | 172 |
| 18 | 64 | 87 |
| 19 | 308 | 282 |
| 20 | 149 | 141 |
| 21 | 430 | 420 |
| 22 | 197 | 185 |
| 23 | 335 | 312 |
| 24 | 363 | 353 |
| 25 | 348 | 345 |
| 26 | 145 | 151 |
| 27 | 149 | 143 |
| 28 | 156 | 121 |
| 29 | 373 | 344 |
| 30 | 179 | 172 |
| 31 | 348 | 315 |
| 32 | 376 | 314 |
| 33 | 375 | 347 |
| 34 | 216 | 167 |
| 35 | 119 | 100 |
| 36 | 408 | 408 |
| 37 | 177 | 229 |
| 38 | 193 | 214 |
| 39 | 195 | 179 |
| 40 | 218 | 220 |
| 41 | 389 | 396 |
| 42 | 397 | 391 |
| 43 | 331 | 328 |
| 44(a) | 188 | 180 |
| (b) | 137 | 124 |
| (c) | 83 | 48 |
| 45 | 278 | 248 |
| 46 | 226 | 163 |
| 47 | 159 | 142 |
| 48 | 175 | 157 |
| 49 | 380 | 366 |
| 50 | 251 | 239 |
| 51 | 128 | 119 |
| 52 | 276 | 267 |
| 53 | 150 | 138 |
| 54 | 355 | 337 |
| 55 | 236 | 238 |
| 56(a) | 133 | 130 |
| (b) | 216 | 129 |
| (c) | 137 | 99 |
| 57 | 91 | 81 |
| 58 | 159 | 126 |
| 59 | 162 | 147 |
| 60 | 261 | 255 |
| 61 | 176 | 179 |
| 62 | 190 | 199 |
| 63 | 81 | 87 |
| 64 | 24 | 52 |
| 65 | 34 | 27 |
| 66(a) | 23 | 24 |
| (b) | 11 | 19 |
| 67 | 232 | 224 |
| 68 | 157 | 137 |
| 69 | 184 | 147 |
| 70 | 53 | 43 |
| 71 | 221 | 220 |
| | 18546 | 17763 |

Picture Test

| Question | Boys (450) | Girls (450) |
|---|---|---|
| 1 | 431 | 437 |
| 2 | 416 | 413 |
| 3 | 399 | 392 |
| 4 | 386 | 375 |
| 5 | 402 | 398 |
| 6 | 389 | 375 |
| 7 | 179 | 184 |
| 8 | 268 | 251 |
| 9 | 205 | 198 |
| | 3075 | 3023 |

Table 46. Score-scatters of M.H.T. 12.

Verbal Test

| Score | Frequencies Boys | Girls | Total |
|---|---|---|---|
| 0 - 9 | 3568 | 2755 | 6323 |
| 10 - 19 | 5104 | 4998 | 10102 |
| 20 - 29 | 7263 | 7620 | 14883 |
| 30 - 39 | 10123 | 10766 | 20889 |
| 40 - 49 | 10064 | 10282 | 20346 |
| 50 - 59 | 6207 | 5439 | 11646 |
| 60 - 69 | 1800 | 1376 | 3176 |
| 70 - 76 | 81 | 52 | 133 |
| | 44210 | 43288 | 87498 |

Picture Test

| Score | Boys | Girls | Total |
|---|---|---|---|
| 0 | 923 | 721 | 1644 |
| 1 | 890 | 749 | 1639 |
| 2 | 1043 | 1044 | 2087 |
| 3 | 1322 | 1350 | 2672 |
| 4 | 1403 | 1616 | 3019 |
| 5 | 2437 | 2746 | 5183 |
| 6 | 6792 | 6681 | 13473 |
| 7 | 9530 | 9089 | 18619 |
| 8 | 10606 | 9895 | 20501 |
| 9 | 9264 | 9397 | 18661 |
| | 44210 | 43288 | 87498 |

The score-scatters for the total population are graphed overleaf. As has been shown in chapter 4 of Part 1, the answer-pattern obtained from the boys was essentially the same as that obtained from the girls, so that either may be taken as representative. This was true of both the tests 12v and 12p. These answer-patterns were graphed on pages 46 and 49.

After examination of the results given, the Scottish Council for Research in Education reissued this test in a revised form as the "1932 Mental Survey Test". In this form the items of the verbal part have been rearranged in ascending order of difficulty wherever possible, and minor changes have been made in the phrasing of some of the items. While the order of the pictorial items is left unchanged, an improvement has been effected by the printing of these items upside down, to prevent the very weak pupils from turning back to this page for fresh attempts when they should be tackling the verbal items.

When sufficient data from this form of the test become available, it will be very interesting to find whether the change of order of the items in 12v has affected the answer-pattern to any appreciable extent.

Figure 32. Score-scatter of M.H.T. 12v.
( 87498 candidates )
— 44210 boys
— 43288 girls

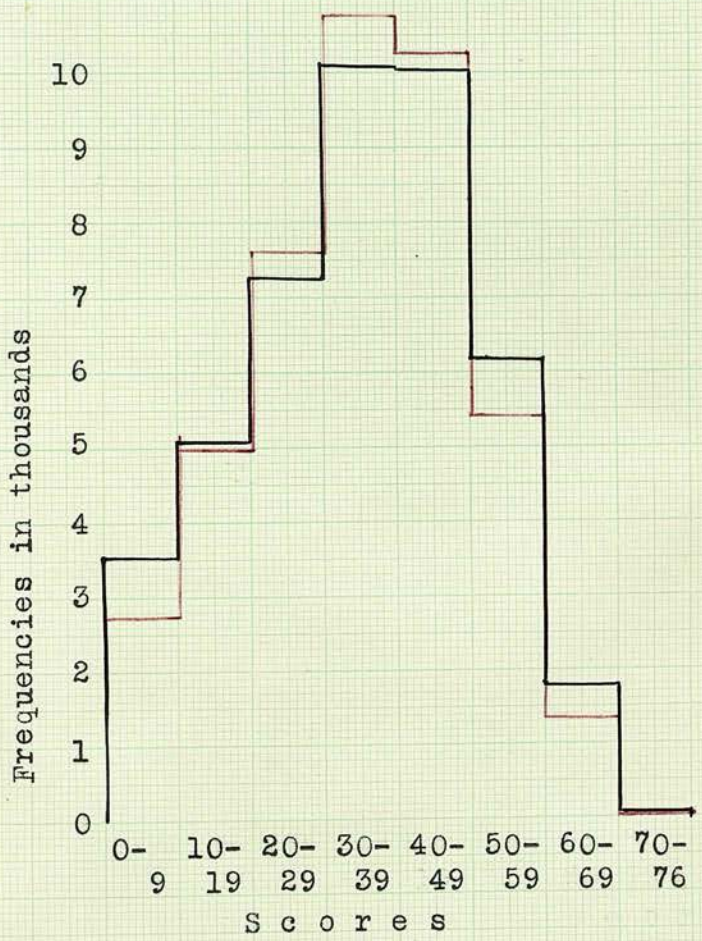
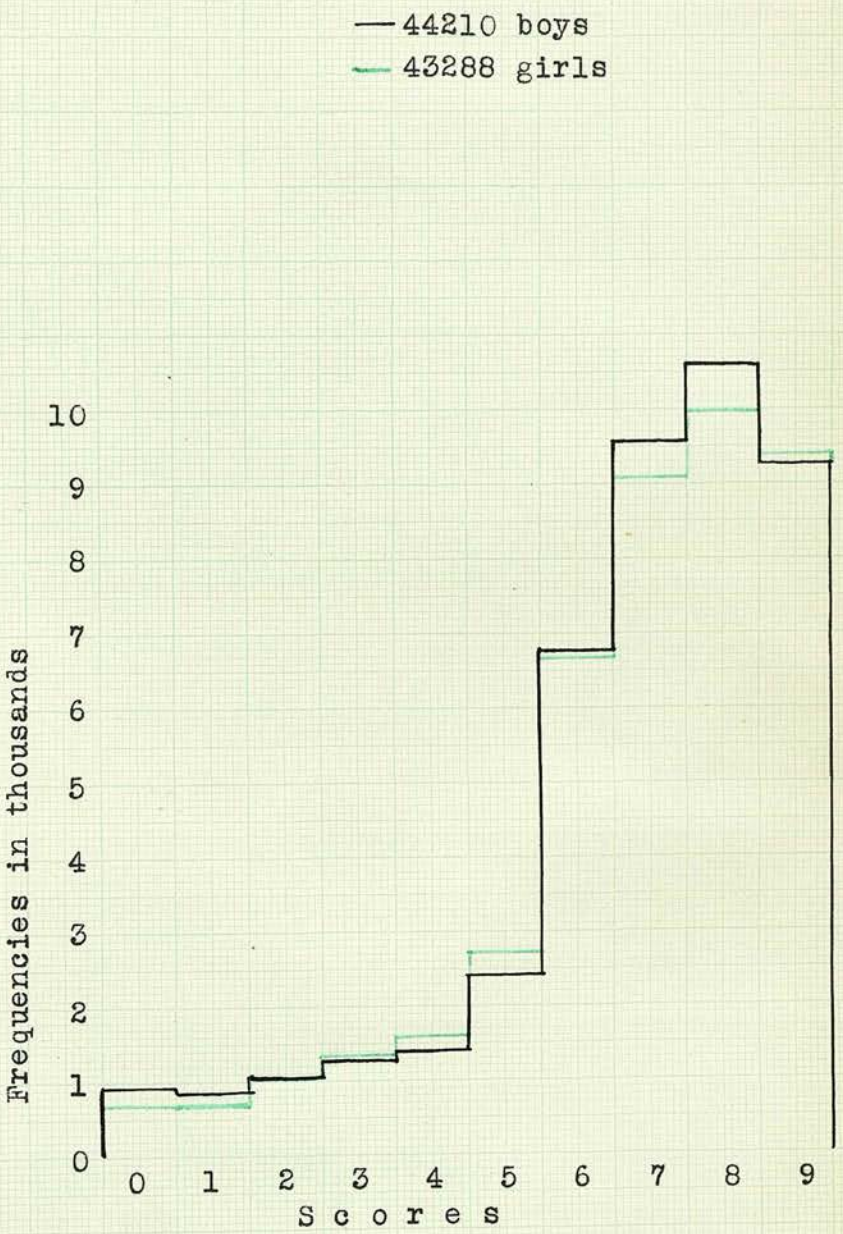


Figure 33. Score-scatter of M.H.T. 12p .



# PART THREE

This part contains two published papers on the subject.

1. Answer-pattern and Score-scatter in Tests and Examinations.
   British Journal of Psychology, XXII, 73-86, 1931.
2. Answer-pattern and Score-scatter in Tests and Examinations.
   British Journal of Psychology, XXVI, 301-308, 1936.

NOTE. Some discrepancies may be observed between the conclusions and numerical results of these papers and those given in Part One. The alterations are due partly to the adoption in Part One of better statistical methods, and partly to the use in that part of a greater volume of evidence.

[From THE BRITISH JOURNAL OF PSYCHOLOGY (GENERAL SECTION), Vol. XXII, Part 1, July, 1931.]



# ANSWER-PATTERN AND SCORE-SCATTER IN TESTS AND EXAMINATIONS[1].

By DAVID A. WALKER.

(*From the Department of Education, University of Edinburgh.*)



## I. Introduction.

It is a well-known fact that it lies in the power of an examiner to obtain, within certain limits, whatever type of score-scatter he desires, and that he does this by a suitable choice of questions. The methods of choice are usually of the rule-of-thumb type, experience or intuition being the guide. In this paper an attempt is made to investigate the theoretical basis of the preordaining of the score-scatter, and an account is given of an experiment in which tests were designed on this theoretical basis to produce score-scatters of given types. As will be seen in the course of the paper, the problem is really a dual one: "(i) How, and (ii) how far, do the examiner's plans preordain the score-scatter?"

The examiner has control over the test and its details, such as the number of scoring points, difficulty of questions, time allowance, and so on. A factor over which he has no control, but which has a great effect on the type of score-scatter obtained, is the character of the group of testees. As being outside the examiner's control, this factor is largely ignored in the present investigation.

The type of test to be considered here consists of independent items, any one of which may be answered correctly independently of other questions. The difficulty of the individual question in relation to the

[1] Condensed from a thesis submitted in part fulfilment of the requirements for the degree of Bachelor of Education.

difficulty of the other questions is then the main tool of the examiner in preordaining the score-scatter. In any individual test there may be other factors present, such as a linking of certain questions in a degree of interdependence, which will affect the score-scatter; but such factors which are peculiar to the single test in question are here ignored.

## II. Answer-pattern and score-scatter.

The varying degree of difficulty of the questions will be reflected in the different frequencies with which they are correctly answered; an easier question will be answered correctly more frequently than a harder question, and, if certain conditions later to be stated are fulfilled, this is the only test of degree of difficulty. If question 1 is answered correctly $n_1$ times, question 2 $n_2$ times, and so on, then these questions will be placed in order of difficulty when their $n$'s are in order of magnitude. The table of values of $n$ for the questions of a test, placed in order of magnitude, is called the 'answer-pattern' of that test. A test has not a unique answer-pattern, for the pattern depends not only on the difficulties of the items, but also on the character of the group of testees. The characteristic answer-pattern may be taken as that produced by a 'normal' selection, a fair sample, of the population.

A useful extension of this nomenclature is to let $n_0$ represent the number of candidates.

The capital letters $N_x$ are used to denote the numbers of candidates scoring $x$ marks exactly. It is obvious that the $n$'s and the $N$'s are related in some ways. It can readily be seen that

$$\sum_0^m N_x = n_0,$$

and that if $m$ is the number of items in the test

$$\sum_1^m xN_x = \sum_1^m n_x.$$

In all subsequent work it is assumed that the questions have been placed in order of difficulty so that

$$n_0 \geqslant n_1 \geqslant n_2 \geqslant n_3 \ldots \geqslant n_m.$$

## III. 'Hig' and 'Unig.'

A score of exactly $x$ in a test may be made up in many different ways. The most probable composition is that obtained by answering the $x$ easiest questions, but owing to individual differences not every score $x$ is

actually obtained in this unique manner. An element of 'higgledy-piggledyness' enters into the composition of all save zero, and perfect, scores, for which Prof. Godfrey Thomson has suggested the name 'hig,' and for its converse the term 'unig.' These terms are used throughout this paper. By a test being unig we mean that each score $x$ is composed of correct answers to the $x$ easiest questions, and therefore of correct answers to no other questions. Hig implies a departure from this composition. Note that it is not sufficient for our purposes to define unig by stipulating that every score $x$ is identical in composition—there must be added the condition that it is composed of the $x$ easiest items; in other words the score $x + 1$ always comprises the $x$ items of the score $x$, and one more.

Now if hig is absent, that is each score is unig, it is easy to show that an *exact* relationship exists between the $n$'s of the answer-pattern and the $N$'s of the score-scatter. Symbolically

$$\left.\begin{aligned} n_0 &= N_0 + N_1 + N_2 + N_3 + \ldots + N_m \\ n_1 &= \quad N_1 + N_2 + N_3 + \ldots + N_m \\ n_2 &= \quad\quad N_2 + N_3 + \ldots + N_m \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ n_m &= \quad\quad\quad\quad\quad\quad N_m \end{aligned}\right\} \quad \ldots\ldots(A),$$

or in another form

$$\left.\begin{aligned} n_0 - n_1 &= N_0 \\ n_1 - n_2 &= N_1 \\ n_2 - n_3 &= N_2 \\ &\cdots\cdots\cdots\cdots \\ n_{m-1} - n_m &= N_{m-1} \\ n_m &= N_m \end{aligned}\right\} \quad \ldots\ldots(B).$$

*These equations show that when hig is absent a given answer-pattern completely determines the score-scatter.* The relationship may be very neatly illustrated graphically; the difference of adjacent ordinates on the $n$ or answer-pattern graph being equal to the corresponding ordinates of the $N$ or score-scatter graph. In the limit, when there are innumerable items, equations (A) become

$$n_x = \int_x^\infty N dx,$$

equations (B) become $$-\frac{dn}{dx} = N_x.$$

From the above equations and their graphical representations a rather important conclusion may be drawn. The second set of equations

may be interpreted as "the ordinates of the score-scatter curve are given by the *slope* of the answer-pattern." Thus to separate out candidates at the top by skewing the score-scatter positively, with the tail to the right, it is necessary to provide an answer-pattern falling sharply at first, and then falling slowly over the upper half of the items. That is, a relatively large number of difficult questions must be present, but *it is not necessary for any of the questions to be outstandingly difficult.* The spacing out of the candidates occurs as a result of the *differences* between adjacent questions being slight, not as a result of the absolute difficulty of any question. Two test papers of exactly the same range of difficulty may, by a proper choice of intermediate questions, skew the score-scatter in opposite directions.

### IV. Occurrence of hig.

The above exact statements refer to cases where hig is absent. But in actual practice, hig enters in varying degree into the composition of every test result. Due to individual variations in preference or ability for certain questions, all scores of $x$ are not made up uniquely of answers to the $x$ easiest items. Thus there is destroyed, in greater or less degree, the exactness of the relation established between answer-pattern and score-scatter, according as the amount of hig present is large or small.

The amount of hig present depends on various factors. One such, excluded here, is the linking of questions so that a correct answer to one is only possible after a correct answer to the preceding one. A second, having much the same effect, is the enlarging of the steps of difficulty between the questions. Then a correct answer to a difficult question will very probably be accompanied by a correct answer to the preceding question, since it is so much easier, and so on. On the other hand, a test whose questions are all of the same difficulty will tend to show a great degree of hig in the answering. These two statements will be proved mathematically below. Lastly, the injunction "Begin at the beginning and go straight through" usually found in group tests, will tend to decrease the amount of hig. This last factor is psychological and cannot easily be quantitatively investigated. The second factor mentioned deserves a more detailed investigation.

Since large steps of difficulty between questions will be shown in a steep answer-pattern graph, we may designate such a test as 'steep,' and the opposite type of test as 'flat.' This is not an exact definition, but it serves a useful purpose here. We shall now prove that 'steepness' decreases hig, while 'flatness' makes it a maximum.

## V. Relation of hig and steepness.

Let the $m$ questions of a test be arranged in order of descending difficulty and then numbered 1, 2, 3..., $m$. Let the *a priori* probability of answering correctly question $a$ be $p_a$, and so on. Then the probability of scoring exactly $x$ in unig fashion is

$$p_1p_2p_3\ldots p_x\,(1-p_{x+1})\,(1-p_{x+2})\ldots(1-p_m).$$

The probability that any given score $x$ has been compiled in unig fashion is then

$$\frac{p_1p_2p_3\ldots p_x\,(1-p_{x+1})\,(1-p_{x+2})\ldots(1-p_m)}{\Sigma\, p_ap_bp_c\ldots p_e\,(1-p_s)\,(1-p_t)\ldots(1-p_w)},$$

where each term of the denominator is composed of $x$ $p$'s and the corresponding $m-x$ $(1-p)$'s multiplied together, and $\Sigma$ denotes the sum of all the ${}_mC_x$ such terms, including that already appearing in the numerator. The numerator is the chance of getting $x$ marks in unig fashion, the denominator is the sum of the chances of getting $x$ marks in any fashion whatever, and their ratio is the chance that a given score $x$ is actually unig.

Now $p_a$ is evidently proportional to $n_a$; substituting and eliminating the constant of proportionality the probability, of any score being unig is seen to be

$$\frac{n_1n_2n_3\ldots n_x\,(n_0-n_{x+1})\,(n_0-n_{x+2})\ldots(n_0-n_m)}{\Sigma\, n_an_bn_c\ldots n_e\,(n_0-n_s)\,(n_0-n_t)\ldots(n_0-n_w)},\ \text{or } u_x.$$

The probability that a whole test is unig therefore is

$$\Pi=\prod_{x=1}^{x=m}u_x,$$

and the probability of hig in a test equals $1-\Pi$.

Now in a flat test $n_1=n_2=n_3=\ldots=n_m$. The probability of hig therefore becomes

$$1-\Pi\,\frac{1}{{}_mC_x},$$

which rapidly approaches 1 or certainty as $m$ increases. (For $m$ equal to 3, $1-\Pi$ equals 0·89.) That is, for a flat test the probability of hig is a maximum.

Similarly, in a steep test, it can be shown that the above product $\Pi$ tends towards unity and therefore that the probability of hig tends to zero.

*Therefore a steep test reduces the amount of hig, and a flat test gives the maximum amount of hig.*

## VI. Coefficient of hig $h$.

Once the results of a test are known, the amount of incidence of hig can be measured. Hig represents a deviation from the exactness of the relations (B), and may therefore be measured by the amount of such deviation, or better by the sum of the squares of the deviations. Consider then

$$\sum_{0}^{m} (n_x - n_{x+1} - N_x)^2.$$

When hig is absent this equals zero, and conversely, when this is zero hig is absent. To convert this measure into a coefficient, divide by the maximum amount of hig. This, as shown above, occurs when

$$n_1 = n_2 = n_3 \ldots = n_m = n, \text{ say.}$$

Then $$\sum_{0}^{m} (n_x - n_{x+1} - N_x)^2$$

becomes $$\sum_{0}^{m} N_x^2 + n^2 + (n_0 - n)^2 - 2nN_m - 2\,(n_0 - n)\,N_0.$$

Using this as denominator, let $h$ the coefficient of hig be defined[1] as

$$h = \frac{\sum_{0}^{m} (n_x - n_{x+1} - N_x)^2}{\sum_{0}^{m} N_x^2 + n^2 + (n_0 - n)^2 - 2nN_m - 2(n_0 - n)\,N_0}.$$

In the majority of tests discussed $N_0 = N_m = 0$, and

$$h = \frac{\sum_{0}^{m} (n_x - n_{x+1} - N_x)^2}{\sum_{1}^{m} N_x^2 + n^2 + (n_0 - n)^2}.$$

It may clarify the above symbolism and reasoning if one fictitious example be shown, in which the *a priori* probability of hig and the actual amount of hig present are calculated. The table shows the score sheet of the twelve candidates who sit a test comprising three questions (for $m > 3$ the calculation of *a priori* hig becomes very heavy). A cross in the square represents a correct answer by that candidate to the corresponding question.

[1] Since the completion of this work and during the task of condensing the thesis into the present form, it has occurred to me that Pearson's measure of goodness of fit of the differential of the answer-pattern to the actual score-scatter would be a measure of 'hig,' and would be closely connected with the coefficient here employed; and I hope to explore this avenue.

| | Question 1 | 2 | 3 | |
|---|---|---|---|---|
| A | × | × | × | 3 |
| B | × | × | × | 3 |
| C | × | × | × | 3 |
| D | × | × | | 2 |
| E | × | | × | 2 |
| F | | × | × | 2 |
| G | × | × | | 2 |
| H | × | | × | 2 |
| I | × | × | | 2 |
| J | × | | | 1 |
| K | | × | | 1 |
| L | | | × | 1 |
| | 9 | 8 | 7 | |

Answer-pattern

| $x$ | 0 | 1 | 2 | 3 | Sum |
|---|---|---|---|---|---|
| $N_x$ | 0 | 3 | 6 | 3 | 12 |
| $N_x^2$ | 0 | 9 | 36 | 9 | 54 |
| $n_x$ | 12 | 9 | 8 | 7 | |
| $n_x - n_{x+1}$ | 3 | 1 | 1 | 7 | |
| $\lvert n_x - n_{x+1} - N_x \rvert$ | 3 | 2 | 5 | 4 | |
| $(n_x - n_{x+1} - N_x)^2$ | 9 | 4 | 25 | 16 | 54 |

$n_0 = 12.$
$n = 24/3 = 8.$
$h = 54/86 = 0{\cdot}63.$

The *a priori* probability of hig by the former formula is $1 - 0{\cdot}22 = 0{\cdot}78$. Taking the steepness as being roughly proportional to $n_1 - n_3$, this test had steepness 2. A further four similar examples, plus the above, yielded the following table:

| Test | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Relative steepness | 9 | 5 | 4 | 2 | 0 |
| *A priori* hig | 0·10 | 0·59 | 0·63 | 0·78 | 0·89 |
| Coefficient $h$ | 0 | 0·19 | 0·35 | 0·63 | 1·0 |

The positive correlation between *a priori* hig, the coefficient $h$, and the inverse of the steepness is obvious. Naturally the relations cannot be exact.

The conclusions so far arrived at may be conveniently summarized here before the experiment proper is described:

(i) The type of score-scatter, with little or no hig present, is determined by the shape of the answer-pattern.

(ii) The closeness with which the score-scatter and answer-pattern are related depends on the amount of hig present.

(iii) The amount of hig present depends on the steepness of the answer-pattern.

(iv) The amount of hig present can be measured.

These theoretical conclusions were tested in the experiment now to be described.

## VII. The experiment.

As opposed to the usual school subject test, the typical group intelligence test conforms very well to the conditions for preordaining the score-scatter. It consists of a number of independent items ranging in difficulty from very easy to very difficult, *i.e.* it is steep. The injunction "Begin at the beginning and go straight through" also helps to minimize the amount of hig present. One such group intelligence test (Moray House Test 9) consisting of 94 items was chosen for the experiment. It had been independently standardized, and had also been used in two areas in England. For our present purpose a special preliminary trial on 30 children was made, and the answer-pattern proved to be steep as expected, the number of correct answers to each item ranging from 29 to zero. There were no interdependent items in this test.

The experiment proper consisted in designing, by choice from the items of the above test, three tests to give certain types of score-scatter, and then examining the answer-patterns and score-scatters actually obtained. The first test, test $A$, was to be designed to give maximum hig, and therefore maximum freedom to the score-scatter. Tests $B$ and $C$ were to be designed to produce score-scatters skewed positively and negatively respectively. For them it was therefore essential to have steep answer-patterns, and moreover answer-patterns of given shapes, so that their differences should provide the required skew curves.

Now Moray House Test 9 had been shown to be steep, and at the same time reliable data were procurable from which the answer-pattern could be obtained, viz. results from 800 children in an English borough. From this test therefore it was decided to make three tests $A$, $B$, and $C$, each of 15 items, corresponding to the above description.

The work of finding the answer-pattern was facilitated by the use of mechanical counters mounted on a stand. The 800 papers were treated 100 at a time, and after the second hundred had been finished it was found that the correlation between the answer-patterns of these first two hundreds was, by the Footrule, 0·997 with a P.E. of 0·0004, and therefore it was concluded that the 200 (really 202) papers formed a sufficient sample. The answer-pattern of these 202 papers was practically

a straight line (see Fig. 1). From such an answer-pattern with a low degree of hig there should result a score-scatter consisting of a straight line parallel to the $x$-axis, *i.e.* every score should have about the same number of candidates. Now the degree of hig was very low ($h = 0{\cdot}043$); yet the score-scatter was the normal curve usually obtained in intelligence tests, so that the important conclusion can be drawn that *the normal or*

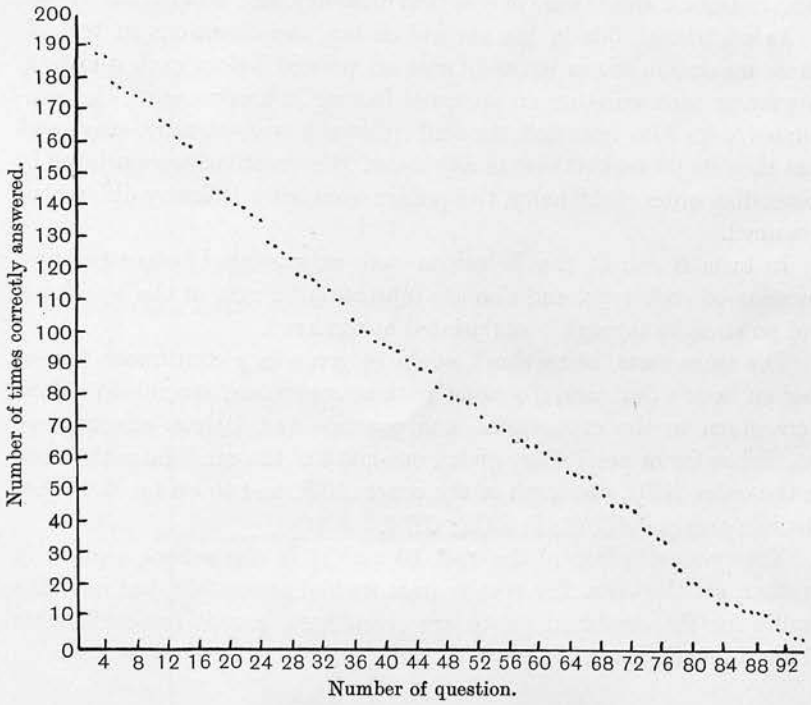



Fig. 1. Answer-pattern of Moray House Test 9. (202 candidates.)

*Gaussian character of the distribution of mental ages found with intelligence tests of the type of Moray House Test* 9 *is not caused by, but occurs rather in spite of, the nature of the answer-pattern coupled with a very low degree of hig.*

From the above data the three tests $A$, $B$, and $C$ were now constructed. Test $A$ was designed to give a maximum amount of hig. Its 15 items were therefore chosen to be of as uniform difficulty as possible, a medium[1] degree of difficulty being chosen.

[1] Medium, that is, for groups of the ability level of the former Moray House Test 9 sample. Unfortunately it did not prove medium for the group tested.

Test *B* was designed to give the minimum amount of hig, and a score-scatter skewed positively, *i.e.* with the tail to the right, or so as to give a median score of 5 and space out the candidates at the top. Its items therefore were chosen to increase rapidly in difficulty to question 10, then slowly to question 15. Test *C*, on the other hand, had items increasing slowly in difficulty up to question 5, and then quickly up to 15. The first items in tests *B* and *C* were of identical difficulty, and also the last items.

As additional aids in the control of hig, the directions in test *A*, where maximum hig is wanted, were all printed before each question, so putting each question on an equal footing in this respect. The candidates were also informed that all questions were equally easy, and that they might tackle them in any order. The questions were printed in descending order of difficulty, to equalize what little difficulty differences remained.

In tests *B* and *C*, the directions were only printed before the first question of each type, and also the injunction "Begin at the beginning and go straight through" was printed at the head.

The three tests, being short, could be given in a continuous test of half an hour's duration. To equalize time conditions, special directions were given to the supervisors, and practice and fatigue effects were avoided as far as possible by giving one-sixth of the candidates the tests in the order *ABC*, one-sixth in the order *ACB*, and so on for the other possible orders *BAC*, *CAB*, *BCA*, *CBA*.

Ninety-six children of the ages 10 and 11 in one school, and 70 in another, sat the tests. The results were treated separately, but only the results for the combined group are given here, except for coefficients of hig, tabulated separately.

## VIII. Results.

*Answer-patterns obtained.* These are shown in Fig. 2. It will be seen that, while some items seem to have altered their positions slightly in the order of difficulty, the essential features of each test's answer-pattern have been preserved.

*Hig.* Test *A*, it will be remembered, was designed to give maximum, tests *B* and *C* minimum hig. The actual coefficients of hig obtained are shown in the table:

| Test | *A* | *B* | *C* | |
|---|---|---|---|---|
| 1st group data | 0·35 | 0·09 | 0·045 | 96 cases |
| 2nd group data | 0·54 | 0·12 | 0·13 | 70 ,, |
| Total data | 0·40 | 0·092 | 0·065 | 166 ,, |

The disparity between $h_A$ and $h_B$ or $h_C$ is obvious.

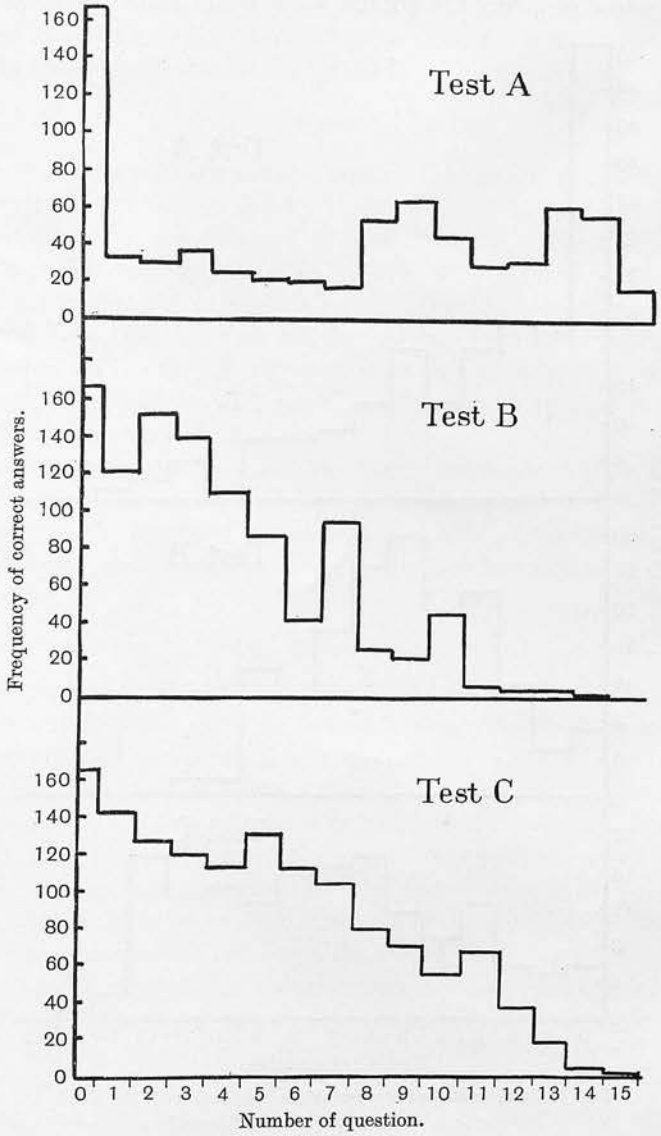



Fig. 2. Histograms of answer-patterns, *i.e.* number of times each question is correctly answered, obtained with the three tests A, B and C. (166 candidates.)

*Score-scatter* (see Fig. 3). The most noticeable feature, at first, is that test *A*, instead of giving the normal curve one would expect from the

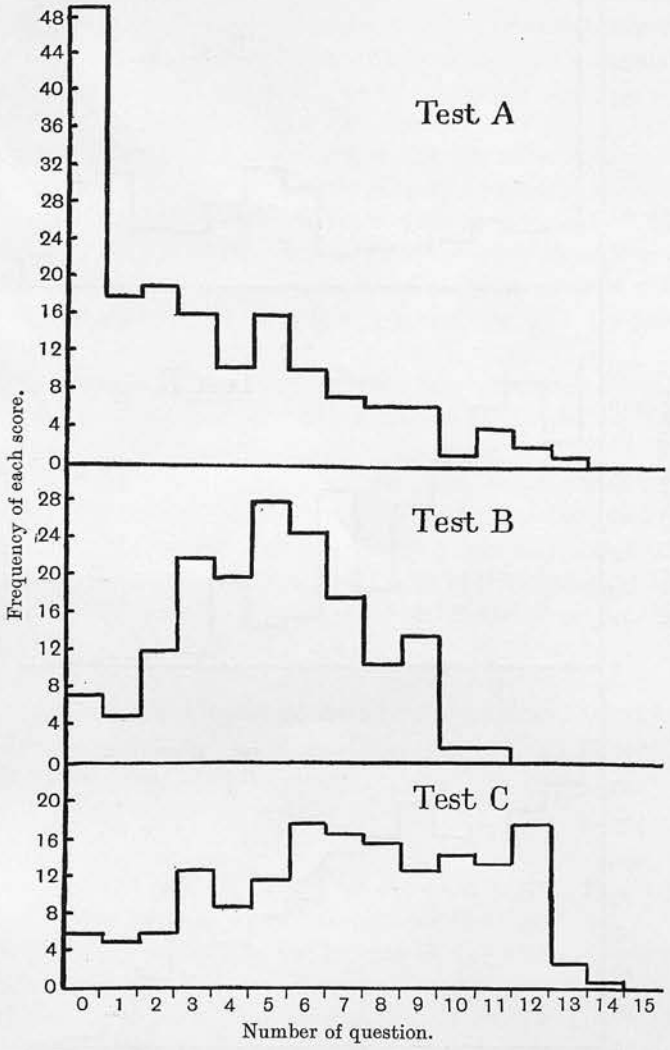



Fig. 3. Histograms of score-scatters obtained with the three tests A, B and C. (166 candidates.)

level nature of the answer-pattern, has yielded such a large number of zero scores. This is due to the ability level of the population tested being lower than that of the sample which sat Moray House Test 9. The

raising of the level of difficulty of test $A$ will tend to skew the score-scatter positively, and this is what has happened.

Test $B$ shows a median score of 5 as required and is skewed positively, so as to open out the candidates at the top. The formula

$$\text{skewness} = \frac{\mu_3}{\sigma^3}.$$

where $\mu_3$ = third moment and $\sigma$ = standard deviation, gives here a value $S = 0{\cdot}01$, positive though slight.

Test $C$, which was designed to give a negative skewness, so as to space out the candidates at the bottom, does so very satisfactorily, having a coefficient of skewness of $-0{\cdot}24$. The effect is rather clouded in the total data, but the first set showed the separation of the poorer candidates very well. Of 96 candidates, 2 scored zero, 1 scored 1, 3 scored 2, 8 scored 3 and 2 scored 4, the other 80 candidates being in the region of scores 5 to 15.

Both the purposes for which the tests were designed have therefore been fulfilled, the amount of hig has been reduced to a very small amount, and the two tests ($B$ and $C$) have yielded, from the same group of candidates, score-scatters skewed in opposite directions.

## IX. CONCLUSIONS.

The following conclusions apply only to those tests which are composed of independent items. In the above paper they have been derived theoretically and tested practically.

1. The factors which predetermine the character of the score-scatter are the individual and relative difficulties of the questions, the graph or table of which has been called the answer-pattern. Each type of answer-pattern *tends* to produce its own type of score-scatter, the curve of the latter tending to be the differential of the curve of the former.

2. The factor which influences the extent of such predetermination is what has been called 'hig' (higgledypiggledyness of answering). Complete absence of hig leads to complete dependence of score-scatter on the answer-pattern. Experiment showed, however, that even a small amount of hig allows the 'normal' nature of score-scatters latitude to assert itself, so that with intelligence tests the normal distribution of mental ages found is not caused by, but occurs in spite of, the nature of the answer-pattern.

3. The incidence of hig depends on the steepness of the answer-pattern of the test, so that from the answer-pattern it is possible to pre-

dict not only the most probable type of score-scatter but also the degree of hig likely to be present, and therefore the accuracy of the prediction of the score-scatter.

4. An exact measure of the incidence of hig in any test whose results are known is provided by the coefficient $h$.

5. It is rather the differences of difficulty between the questions, than the absolute difficulty of the hardest, or the easiest, question, which determines whether a test will open out the top, or the bottom, candidates.

(*Manuscript received* 19 *November*, 1930.)

[From THE BRITISH JOURNAL OF PSYCHOLOGY (GENERAL SECTION), Vol. XXVI. Part 3, January, 1936.]


# ANSWER-PATTERN AND SCORE-SCATTER IN TESTS AND EXAMINATIONS

By DAVID A. WALKER.

(*From the Department of Education, Edinburgh University.*)



## I. Introduction.

In a previous paper in the *British Journal of Psychology*,[1] a report was made of an investigation into the factors influencing the score-scatters made in tests and examinations. It is a well-known fact that an experienced examiner, by a suitable choice of questions and style of paper, can obtain, within limits, the type of score-scatter he desires from a given group of candidates. In the above paper an attempt was made to trace the underlying factors which produce the desired score-scatter. In the present paper the investigation is carried further, the experimental evidence especially having been considerably amplified.

It was shown in the above paper that an important factor influencing the score-scatter was the answer-pattern-differential, a term defined there. It would probably be convenient to recapitulate briefly the points in the derivation of this curve.

A test paper may consist of items all of the same standard of difficulty, or, while preserving the same average difficulty, it may contain items ranging from very difficult to very easy; again, the graduations may be equal or unequal. A convenient way of representing this is to plot against each item $x$ the number of times ($n_x$) it is correctly answered in a

[1] (1931), xxii, 73–86.

given application of the test. The items should be arranged in ascending order of difficulty, *i.e.* $n_1 \geqslant n_2 \geqslant n_3 \geqslant \ldots \geqslant n_m$, where $m$ is the number of items in the test. A useful extension of this nomenclature is to denote the number of candidates by $n_0$. Then the curve or histogram of these $n$'s, *including* $n_0$, is called the answer-pattern of the test.

It may be as well to mention here that a test or examination has not a unique answer-pattern or difficulty level, any more than a given group of candidates has a unique score-scatter which it always produces. An answer-pattern depends upon the candidates as well as on the test items.

From the answer-pattern is derived the important curve or histogram called the answer-pattern-differential. Against each item $x$ (for $x=0, 1, 2, \ldots, m$) is plotted the value of $n_x - n_{x+1}$, the ordinate for the last item being $n_m$. The importance of this curve lies in the fact that in a certain case (when each candidate's score is made up of answers to the easiest items) the score-scatter is identical with it. This was proved in the previous paper; and in the present paper I hope to show that in all the cases examined experimentally, some measure of relation is still there, though in no case was the candidate's score so made up.

## II. Experimental data.

The actual relation between the score-scatter and its corresponding answer-pattern-differential can only be studied by using results of tests in which not only the score-scatter, but also the answer-pattern can be found. Such data are not easily obtained; in the earlier stages of this enquiry they could only be obtained by going over all the papers and counting the number of times each item was correctly answered. Again, not all tests are suitable for this type of investigation; to prevent matters becoming hopelessly complex it is necessary meantime to confine our attention to those tests which are composed of single independent items, each carrying one mark if correctly answered.

Under these circumstances I am greatly indebted to Prof. Thorndike, who sent me a great deal of material from which I was able to extract fairly easily the data I required. I am also indebted to Prof. Thomson, both for his services in procuring Prof. Thorndike's data, and for data from Moray House Tests, in a form suitable for analysis. Without these aids the calculation involved in these problems would have been too much for any one person.

The method of treatment of the data is dealt with under the heading of each aspect of the score-scatter considered.

## III. THE SPREAD OF SCORE-SCATTERS.

A comparatively easy problem is the relation between the spread of a score-scatter as measured by its standard deviation ($\sigma_S$), and the spread of its answer-pattern-differential measured in the same way ($\sigma_{APD}$). It was stated above that the score-scatter is related in some measure to the answer-pattern-differential; the correlation between these two $\sigma$'s may be regarded as a measure of the closeness of that relation in one aspect.

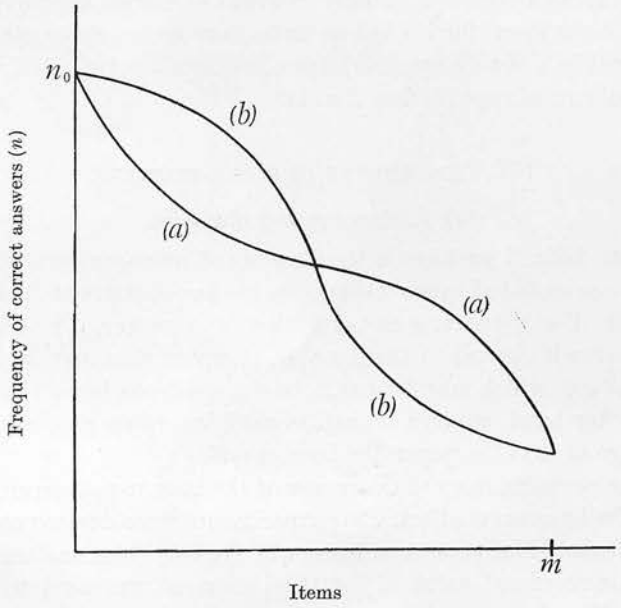



Fig. 1.

From the data sent by Prof. Thorndike were selected the score-sheets of 32 candidates who had worked through 415 items, attempting a fixed number at each sitting. Treating these as 41 tests of 10 items each, attempted by the same 32 candidates all through, we have the data to supply 41 pairs of values of $\sigma_S$ and $\sigma_{APD}$.

The calculation of $\sigma_S$ presents no difficulty, the usual second moment formula being used. In the case of $\sigma_{APD}$, however, there is the difficulty that the answer-pattern-differential in general does not fall under any of Pearson's set of distribution curves, and the value of $\sigma_{APD}$ obtained by the usual method is not strictly a standard deviation. This is not a fatal objection, as the value obtained serves quite well as a measure of

'spread', and no other properties belonging strictly to a standard deviation are used.

The values of $\sigma_S$ and $\sigma_{APD}$ may be grouped and correlated immediately without any difficulty due to heterogeneity, since each of the tests was a 10-item test. The correlation found was $0{\cdot}730 \pm 0{\cdot}050$, a fairly high positive value. So far as the spread of score-scatters is concerned there seems to be a definite relation to the answer-pattern-differential.

An examiner desiring to produce a score-scatter with a high value of $\sigma_S$, and having access to a battery of items of known difficulty, should then select the items for his test so as to have an answer-pattern of the type (*a*); while a test designed to have a small value of $\sigma_S$ should have an answer-pattern of type (*b*) (see Fig. 1).

## IV. The skewing of score-scatters.

### (*a*) *Factors causing skewness.*

A more difficult problem is the skewing of score-scatters. The factor generally suspected of causing skewness in score-scatters is the difficulty of the test. For instance, a recent book on the science of marking says: "If the curve is skewed to the low side, it means that marks have been difficult to get, which may be due...to the questions being too difficult. On the other hand, a curve skewed toward the upper part of the mark scale suggests that the paper has been easy."

On the present theory the skewness of the answer-pattern-differential should also have some effect. Two experiments were devised to test the relative strengths of these two factors in skewing score-scatter.

The standardized value of the third moment was used to measure skewness. The skewness of the score-scatter ($S$) was easily calculated, but as before, the calculation of the skewness of the answer-pattern-differential ($S'$) presented theoretical difficulties, since the curve did not belong to any of the known families of distribution curves. Here again it may be said that $S'$ computed in the same way as $S$ serves to measure the skew tendency of the answer-pattern-differential and no other property peculiar to particular types of distribution is used.

For the purposes of calculation it will be necessary to define more exactly what we mean by difficulty level. If the ratio of the total number of correct answers in a test to the total possible number is $e$, then the difficulty level $d$ may be defined as $1-e$.

The problem may now be stated: "What is the relative influence of each of the factors $S'$ and $d$ in determining $S$?" or, "What is the relative

influence of the shape of the answer-pattern and the average difficulty in determining the skewness of the score-scatter?" The two experiments will be described briefly first, and then the results will be considered.

### (*b*) *Experiment* 1.

The forty-one tests already mentioned were used. For each test the values of $S$, $S'$, and $d$ were calculated. The method of calculation is shown for a typical test in Table I.

Table I. *Test* 4.

Answer-pattern

| Item | *AP* | *APD* | $x$ | $fx$ | | $fx^2$ | $fx^3$ | |
|---|---|---|---|---|---|---|---|---|
| 0 | 32 | 3 | −4 | −12 | | 48 | −192 | |
| 1 | 29 | 2 | −3 | − 6 | | 18 | − 54 | |
| 2 | 27 | 4 | −2 | − 8 | | 16 | − 32 | |
| 3 | 23 | 5 | −1 | − 5 | | 5 | − 5 | |
| 4 | 18 | 6 | 0 | | −31 | | | −283 |
| 5 | 12 | 3 | 1 | 3 | | 3 | 3 | |
| 6 | 9 | 1 | 2 | 2 | | 4 | 8 | |
| 7 | 8 | 1 | 3 | 3 | | 9 | 27 | |
| 8 | 7 | 1 | 4 | 4 | | 16 | 64 | |
| 9 | 6 | 0 | 5 | 0 | | 0 | 0 | |
| 10 | 6 | 6 | 6 | 36 | | 216 | 1296 | |
| Total marks...145 | | ($n_0$)...32 | | | 48 | | | 1398 |
| | | | | +17 | | 335 | +1115 | |

$$d = 1 - \frac{145}{320} = 0{\cdot}547 = \text{measure of average difficulty.}$$

$$m_1 = \frac{17}{32} = +0{\cdot}53,\quad m_2 = \frac{335}{32} = +10{\cdot}47,\quad m_3 = \frac{1115}{32} = +34{\cdot}84.$$

$$\sigma = (m_2 - m_1^2)^{\frac{1}{2}} = 3{\cdot}19.$$

$$S' = \frac{m_3 - 3m_1m_2 + 2m_1^3}{(m_2 - m_1^2)^{\frac{3}{2}}} = +0{\cdot}57 = \text{skewing of answer-pattern-differential.}$$

Score-scatter

| Item | $f$ | $x$ | $fx$ | | $fx^2$ | $fx^3$ | |
|---|---|---|---|---|---|---|---|
| 0 | 1 | −4 | − 4 | | 16 | − 64 | |
| 1 | 1 | −3 | − 3 | | 9 | − 27 | |
| 2 | 1 | −2 | − 2 | | 4 | − 8 | |
| 3 | 10 | −1 | −10 | | 10 | − 10 | |
| 4 | 5 | 0 | | −19 | | | −109 |
| 5 | 4 | 1 | 4 | | 4 | 4 | |
| 6 | 4 | 2 | 8 | | 16 | 32 | |
| 7 | 2 | 3 | 6 | | 18 | 54 | |
| 8 | 2 | 4 | 8 | | 32 | 128 | |
| 9 | 2 | 5 | 10 | | 50 | 250 | |
| 10 | 0 | 6 | 0 | | 0 | 0 | |
| | ($n_0$)...32 | | | 36 | | | 468 |
| | | | +17 | | 159 | +359 | |

Check: $32 \times 4 + 17 = 145$ = total marks scored, as in answer-pattern.
$\sigma = 2{\cdot}17$. $S = +0{\cdot}36$ = skewness of score-scatter.
$d = 0{\cdot}547$, $S' = +0{\cdot}57$, $S = +0{\cdot}36$, $\sigma_{APD} = 3{\cdot}19$, $\sigma_S = 2{\cdot}17$.

From these forty-one sets of values the following correlations were obtained:

$$r_{SS'} = 0{\cdot}617 \pm 0{\cdot}065, \quad r_{Sd} = 0{\cdot}514 \pm 0{\cdot}075, \quad r_{S'd} = 0{\cdot}846 \pm 0{\cdot}030.$$

The regression equation was

$$S = 0{\cdot}64S' - 0{\cdot}03d,$$

and the multiple correlation $R$ of $S$ with the team of $S'$ and $d$ weighted as above was 0·617.

(*c*) *Experiment* 2.

As it is difficult to obtain much data of the above type, where the same candidates have taken part in as many as forty-one tests, a second method of attack on the problem was devised. Here the candidates were not the same for all the tests, but in each case a large number of candidates sat, so that it may be assumed that the distribution of ability was normal. Any skewing of the score-scatter may then be attributed to the joint effects of $S'$ and $d$. The data used were partly selected from those provided by Prof. Thorndike, partly from those published in the previous paper, and partly from results obtained with the Moray House Tests 8, 9, 11, 12 verbal, and 12 pictorial. There were twenty-two tests in all.

The correlations obtained from these were as follows:

$$r_{SS'} = 0{\cdot}754 \pm 0{\cdot}062, \quad r_{Sd} = 0{\cdot}714 \pm 0{\cdot}071, \quad r_{S'd} = 0{\cdot}778 \pm 0{\cdot}057.$$

The regression equation was

$$S = 0{\cdot}48S' + 0{\cdot}31d,$$

and the multiple correlation was $R = 0{\cdot}794$.

Combining the two sets of results, we obtain from the sixty-three tests

$$r_{SS'} = 0{\cdot}627 \pm 0{\cdot}050, \quad r_{Sd} = 0{\cdot}544 \pm 0{\cdot}060, \quad r_{S'd} = 0{\cdot}856 \pm 0{\cdot}024.$$

The regression equation was $S = 0{\cdot}60S' - 0{\cdot}03d$, *i.e.* the shape of the answer-pattern is more important than the average difficulty.

(*d*) *Interpretation of results.*

There are several points to note in considering these results:

(1) The difference between $r_{SS'}$ and $r_{Sd}$ is nowhere large compared to its probable error. On that account any deductions from the results are not very strongly based. The obvious way to rectify this is to repeat the experiments until sufficient data are acquired to reduce the probable errors to small enough dimensions. Unfortunately this is almost im-

possible; the difficulty of obtaining suitable data, and the large amount of calculation involved render it so.

As against this possible objection, it may be noted that in all but one of the many groups of tests afterwards selected from these sixty-three for different purposes, $r_{SS'}$ was greater than $r_{Sd}$.

(2) It is a moot point whether the correlations in which $d$ is concerned should not have had for variable $d^3$, on the ground that $S$ and $S'$ are both of the third dimension. To ensure that non-linear correlations were not being introduced, Blakeman's test of linearity of regression[1] was applied to all the correlations and all were found sufficiently linear.

(3) The high value of $r_{S'd}$ brings rather an unusual feature into the regression equations. This high value might have been expected on the following grounds. In a test with items all of the same difficulty, all being answered correctly $n$ times, say, the answer-pattern becomes the line $y=n$, with an isolated point $y=n_0$ at the beginning. The answer-pattern-differential therefore has ordinates $n_0-n$, 0, 0, ..., 0, $n$. The value of $S'$ for such a distribution can be shown to be $\dfrac{2d-1}{d^{\frac{1}{2}}(1-d)^{\frac{1}{2}}}$, where $d$ is defined as before.

That is, for tests of this type, $S'$ is perfectly correlated with $d$. In other tests the relationship is still shown in some measure. To a certain extent, in measuring $S'$ and $d$ we are measuring the same thing.

What the above results make clear is that, of the two, $S'$ gives the better prediction of the value of $S$ likely to be obtained. In the results of the first experiment, and in the results from the sixty-three tests all together, $S'$ alone gives as reliable a prediction of the probable value of $S$ as does a team of $S'$ and $d$, weighted in the best possible way.

To construct a test which is intended to produce a score-scatter skewed positively, the examiner should therefore work with an answer-pattern falling steeply at first and then flattening out. Conversely a test designed to produce a negatively skewed score-scatter should have an answer-pattern falling gently at first, and increasing in slope to a maximum. Examples of these will be found in the previous paper.

## V. SUMMARY AND CONCLUSIONS.

The results of a test furnish the data for the score-scatter, *i.e.* the distribution of the scores obtained by the candidates. In a similar way there may be constructed the answer-pattern which shows the distribu-

[1] *Biometrika*, IV, 349–50.

tion of the correct answers among the particular items. The relation between the two curves is studied experimentally.

The data used are of two types. In the first case a small number of candidates attempts a large number of items, so that each item is attempted by the same group of testees. In the second case the candidates differ from test to test, but are sufficiently numerous to warrant the assumption that the distribution of ability is normal. The two aspects of the score-scatter especially considered are its standard deviation and its skewness.

From these experimental results it appears that the shape of the answer-pattern is quite a strong factor in determining the shape of the score-scatter, both as regards standard deviation and skewness. It is suggested that examiners may apply this fact in designing tests for special purposes.

In a further paper I hope to discuss the effects of certain methods of strengthening the relationship between the answer-pattern-differential and the score-scatter.

(*Manuscript received* 25 *January*, 1935.)

*The British Journal of Psychology* is issued by the British Psychological Society and published in two Sections, a *General Section* and a *Medical Section* now entitled *The British Journal of Medical Psychology*. Each *Section* appears in Parts quarterly.

Papers for publication in the *General Section* should be sent to Prof. F. C. Bartlett, the Psychological Laboratory, University of Cambridge. Those for publication in *The British Journal of Medical Psychology* should be sent to Dr John Rickman, 11 Kent Terrace, Regent's Park, London, N.W. 1.

Contributors receive twenty-five copies of their papers free. Additional copies may be had at cost price: these should be ordered when the final proof is returned.

The subscription price, payable in advance, is for either *Section* 30*s. net* (post-free); the price of single numbers will depend on the size of each number. Subscriptions may be sent to any Bookseller, or to the Cambridge University Press, Fetter Lane, London, E.C. 4.

## Works consulted.

Note. As far as is known by the author, no work on the subject of this thesis has been published by others. The following list is therefore of a general character. In the course of the calculations, Crelle's "Rechentafeln" were found to be invaluable.


A. C. Aitken. Statistics. ( Manuscript notes )

A. L. Bowley. Elements of Statistics. 1907.

W. Brown and G. H. Thomson. Essentials of Mental Measurement. 1925.

R. A. Fisher. Statistical Methods for Research Workers. 1934.

P. Hartog and E. C. Rhodes. An Examination of Examinations. 1935

K. Holzinger. Statistical Methods. 1928

" Spearman's formula for reliability. Journal of Educational Psychology, 1923. XIV. 302.

T. L. Kelley. Statistical Method. 1923.

" Note on the reliability of a test. Journal of Educational Psychology. 1924. XV. 193.

K. Pearson. Tables for Statisticians and Biometricians.

Scottish Council for Research in Education. The Intelligence of Scottish Children. 1933.

" The 1932 Mental Survey Test.

T. Thomas. The Science of Marking. 1930.

E. L. Thorndike. The Measurement of Intelligence.

C. W. Valentine. The Reliability of Examinations. 1932.

## Index of Definitions and Symbols